UCRL-TH-201207

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Analysis of Hypothetical Promoter Domains of DKFZp564A1164, NPHS1 and HSPOX1 Genes

S. S. Hammond

December 1, 2003

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

ANALYSIS OF HYPOTHETICAL PROMOTER DOMAINS OF *DKFZP564A1164*, *NPHS1* AND *HSPOX1* GENES

UCRL-TH-201207

A University Thesis Presented to the Faculty

of

California State University, Hayward

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Biological Science

By

Sha Hammond

March, 2004

ANALYSIS OF HYPOTHETICAL PROMOTER DOMAINS OF *DKFZP564A1164*, *NPHS1* AND *HSPOX1* GENES

By

Sha Hammond

Approved:	Date:

ACKNOWLEDGMENTS

I would especially like to thank my advisor Lisa Stubbs for giving me the opportunity to do my thesis research in her lab and under her auspicious guidance. I would additionally like to thank my supervisor, Joomyeong Kim, for his technical expertise and advice. Many thanks are due to my thesis committee at California State University Hayward, Stephen Benson and Christopher Baysdorfer, who have taught me so much and whose classes I thoroughly enjoyed. And special thanks to Susan Opp at CSHU for giving me a place to work in her lab during my bachelor studies, she truly started my interest in molecular biology, and I would not be here if it were not for her initial guidance and friendship.

Additional thanks are also owed to the following at Lawrence Livermore National Laboratory: my first supervisor, Laurie Gordon, deserves much appreciation for encouraging me to go back to school while finding ways to help me keep my job at the Lab; Eddie Wehri and Xiaochen Lu for performing the tissue *in situ* hybridizations and pathology; the mapping group, particularly Matt Groza, for helping me find the cosmid clones I needed; and I especially could not have made it with out my officemates Mari Christensen and Eddie Wehri who kept my spirits up day in and day out.

Finally I must thank my mother for her constant enthusiasm and support in my education and career. I am also eternally grateful to my Uncle JB for all his generosity and support. Last but certainly not least, I thank my husband for his love, encouragement and support particularly toward the end.

TABLE OF CONTENTS

FIGURES AND TABLES

APPENDIX

INTRODUCTION

Comparing the human genome to that of a related species, such as mouse, provides a unique perspective for identifying similarities and for finding the genes each sequence may encode. This approach has become a powerful method to identify sequences of specific function, such as gene regulatory activity (Loots *et al.*, 2000). Genome comparison works because the biologically essential features of a genome, such as genes and regulatory elements, are conserved through evolutionary pressure, while the non-essential elements readily acquire mutations and diverge between species. Deleterious mutations that occur within essential DNA are not conserved because they decrease the survival rate of the organism, while advantageous mutations, those that increase or preserve the survival rate of the organism, are conserved. This essential DNA is comprised of the protein coding exons of genes and the regulatory sequences that control their activity (Hardison *et al*, 1997; Hood *et al*, 1993) . The use of comparative sequence alignments is, therefore, an effective tool for providing confirmatory evidence of hypothetical genes by identifying candidate exons and regulatory elements, which can be difficult to ascertain through other predictive methods.

The comparative sequence analysis of human chromosome 19 (HSA19) and related regions in mouse highlighted the positions of more than 1300 genes and associated putative regulatory elements including promoters and enhancers (Dehal *et al.*, 2001). These elements are especially interesting because so little is known about them: for instance only 1871 promoters have been characterized out of the 30,000 total human genes (from the Eukaryotic Promoter Database http://www.epd.isb-sib.ch) (Praz *et al*, 2002). In order to confirm or negate the functional relevance of this large number of predicted regulatory elements, we set out to develop a high throughput pipeline to test for promoter and enhancer function in cultured mammalian cells.

Summary and Significance of the Proposed Research

The focus of this master's thesis project was to develop the basic methods that will underlie a high throughput pipeline, and to use these methods to investigate potential promoter elements in a specific gene-rich region containing loci associated with several human disease loci. The region of focus was a 67kb segment of human chromosome 19q13.1 (segment of Genomic Contig, Genbank accession number NT_011196.11), containing three genes *DKFZp564A1164* (*NLG1*), *NPHS1* and *HSPOX1* (also referred to as *PRODH2*), figure 1. HSA19 was chosen as it has been the focus of my work at Lawrence Livermore National Laboratories (LLNL) and there is a wealth of sequence and experimental data available for analysis of this very gene rich chromosome.

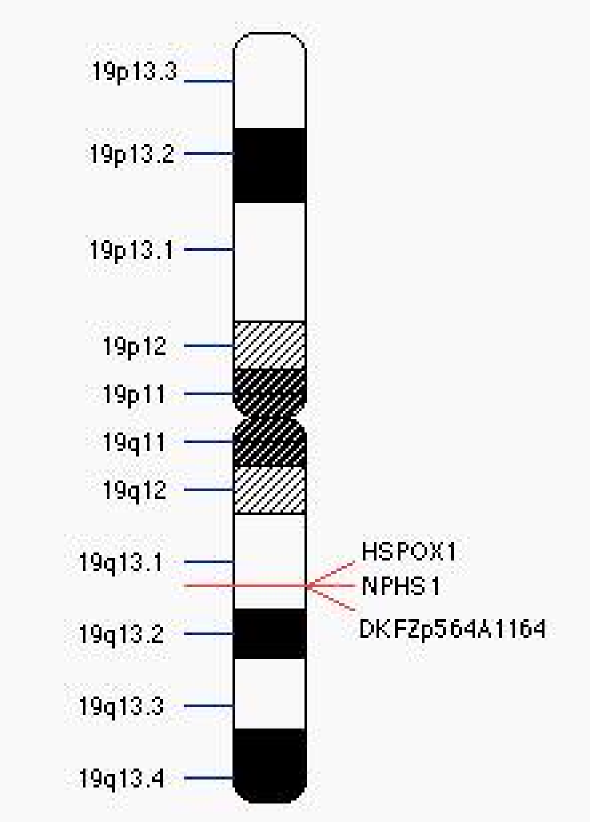


Figure 1: A 67kb segment of human chromosome 19q13.1 (small portion of genomic contig, Genbank accession no.NT_011196.11), containing three genes HSPOX1, NPHS1 and DKFZp564A1164.

Genes of interest

NPHS1, *HSPOX1* and *DKFZp564A1164* were chosen because of their similar expression patterns, and in addition they are part of a larger well-characterized gene-rich region on HSA19q13.1(Locus Link, NT_011196.11). Nephrin, the *NPHS1* gene product, is a 1241-residue putative transmembrane kidney protein of the immunoglobulin family of cell adhesion molecules (Kestila *et al*, 1998) . The disease, congenital nephritic syndrome of the Finnish type, is caused by mutation in the *NPHS1* gene, and exists predominately in Finland (Kestila *et al*, 1998; Lenkkeri *et al.*, 1999). It is characterized by massive proteinuria, detectable in utero by a large placenta and marked edema

(Hallman *et al*, 1956) . The *NPHS1* gene has 29 exons and spans 25.9 kb in length (Genbank accession No., NM_004646).

HSPOX1, also known as *PRODH2*, kidney and liver proline dehydrogenase (oxidase) 2 is located downstream of *NPHS1* and has a very similar expression pattern to that of the *NPHS1* gene. The protein encoded by*HSPOX1* is similar to *PRODH,* proline dehydrogenase (oxidase) 1, a mitochondrial enzyme, which catalyzes the first step in proline catabolism. There is some indication that heterozygous deficiency of *PRODH* on HSA22 may be a cause of isolated hyperprolinemia (Goodman *et al.*, 2000) and schizophrenia susceptibility (Chakravarti, 2002). The known *HSPOX1* gene sequence contains 11 exons and is over 13 kb in length (Genbank accession no.NP_067055). However, the function of the protein encoded by *HSPOX1* has not been determined.

DKFZp564A1164 is a hypothetical protein (Genbank accession no.XP_048303) represented by a cDNA isolated from human fetal brain tissue (AL136654) (Wiemann *et al.*, 2001) and retinoblastoma cells (Genbank accession no.BC007312). As recently as January 2003, Ihalmo *et al.* have described *DKFZp564A1164* as a novel nephrin-like gene (*NLG1*) encoding filtrin, a protein with substantial homology to human nephrin. The known *DKFZp564A1164* coding sequence contains 15 exons and is 10 kb in length. In addition to the full-length form, two alternatively spliced mRNA variants were discovered (Ihalmo *et al.*, 2003). *NPHS1* and *DKFZp564A1164* are transcribed in opposite directions and the distance between the transcription starting points is approximately 5-kb, suggesting that these two genes share a common promoter region and enhancers.

The mouse *Nphs1* gene promoter region has been previously reported and compared to human DNA by sequence alignment (Moeller *et al*, 2000) . The corresponding *NPHS1* gene promoter region in human is conserved in sequence, as highlighted by our percent identity plot (PIP) in figure 2 and VISTA (http://www-gsd.lbl.gov/vista/) alignment in figure 3. However, the precise locations of regulatory elements and start site for transcription of Nphs1 have not been defined.

Sequence comparison tools

The percent identity plot (PIP) is one of the displays available from PipMaker (http://bio.cse.psu.edu), a site for comparing two long DNA sequences to identify conserved segments between species (Schwartz *et al*, 2000) . A PIP shows the position in one sequence of each aligning gap-free segment and plots the degree of similarity between both species as dots or lines (similar to dot plot). For example, PIPMaker can align completed human sequence with homologous mouse DNA even if it is draft sequence, and reveal candidate regulatory elements as highly conserved regions that do not correspond to exons or predicted exons. Positions along the horizontal axis can be labeled with known features such as exons, repetitive elements and CpG islands (Figure 2).

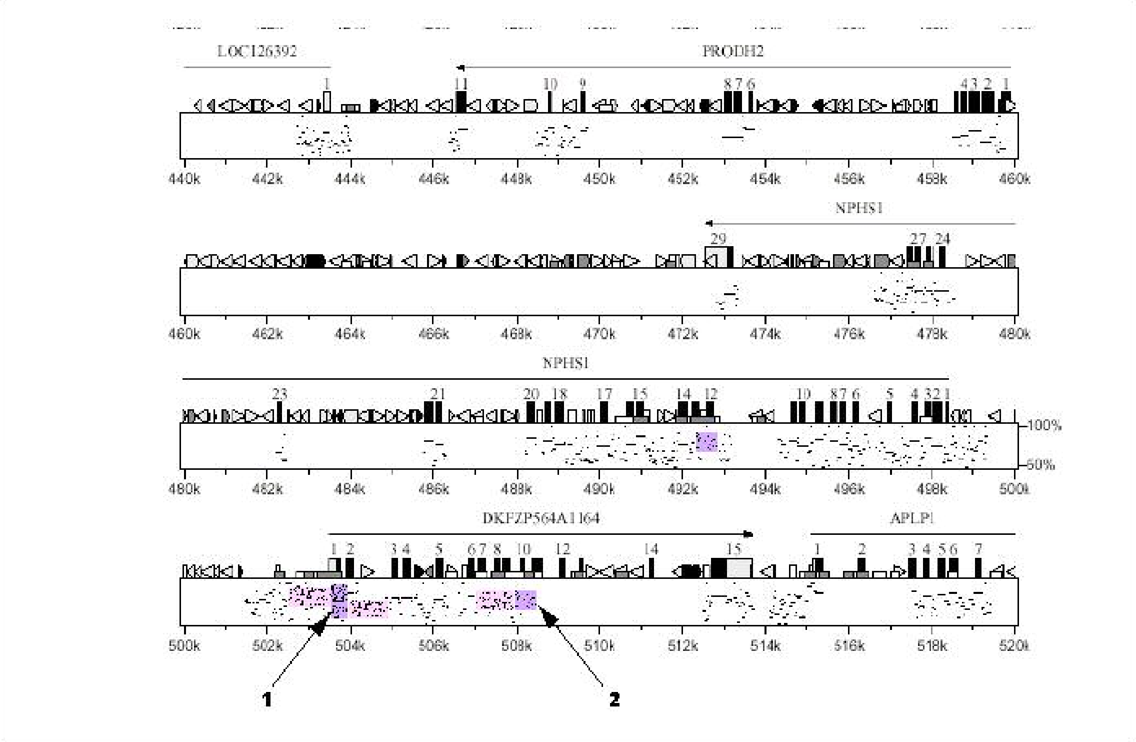


Figure 2: PIP comparing a region in human 19q13.1 and in mouse, highlighting hypothetical promoters (purple) and first exons (pink) predicted by the FirstEF program (Davuluri et al., 2001). Numbers 1 and 2 designate FirstEF promoter predictions for *NPHS1*.

VISTA is a program for visualizing global DNA sequence alignments of arbitrary length. It was designed to visualize long sequence alignments of DNA from two or more species, such as human and mouse, with annotation information (Bray, 2003; Dubchak *et al.*, 2000; Mayor *et al.*, 2000). VISTA is easily configurable, allowing the visualization of alignments of various lengths at different levels of resolution. In figure 3 the x-axis represents base sequences and the y-axis represents percent identity of conserved sequences in the form of graphical peaks. As one can see some segments of DNA are highly conserved whereas other regions are very dissimilar between the human and mouse. Different sequence features such as exons and UTR's are denoted by color coding (Figure 3).

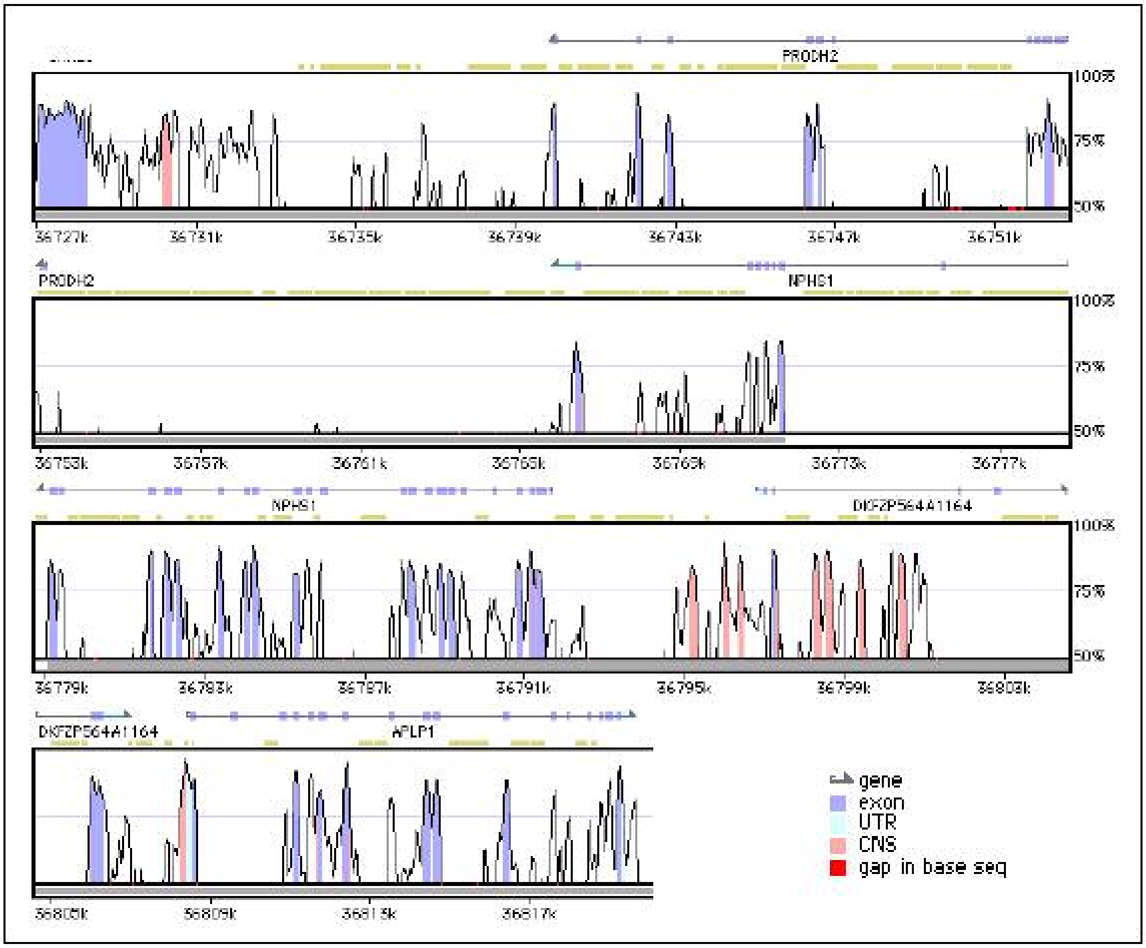


Figure 3: VISTA comparing a region in human 19q13.1 and consensus sequence in mouse, peaks represent conserved sequence.

The human *NPHS1* promoter has not been characterized in laboratory experiments. The hypothetical promoter for this region is 4 to 5 kb upstream of the currently known first exon of the *NPHS1* human gene,suggesting that there is another undiscovered upstream exon for this gene. In fact, many known gene sequences are not complete, in the sense that they include the full protein-coding sequence but do not contain a complete set of non-coding 5’ exons (Davuluri *et al*, 2001) . In addition, a growing body of data suggests that many genes use alternative start sites and promoters in different tissues (Asnagli *et al.*, 2002). Identifying start sites and all promoters used by HSA19 genes is the goal of a larger study in the Stubbs laboratory, and this masters thesis was designed as a focused pilot study to test methods and apply them to analysis of the *NPHS1* gene region.

FirstEF annotation

The positions of promoters (purple) and first exons (pink), which were painted onto figure 2, were predicted by First–exon finder (FirstEF, http://www.cshl.org/mzhanglab), a program developed by M. Zhang and colleagues at Cold Spring Harbor Laboratory (Davuluri *et al.*, 2001). We are working in collaboration with M. Zhang and Zhenyu Xuan (Cold Spring Harbor Laboratory) to confirm the FirstEF predictions in our laboratory using experimental methods.

FirstEF consists of a set of discriminant functions designed to find potential first splice-donor sites and CpG-island related and non-CpG-island related promoter regions. FirstEF decides whether the intermediate region could be a potential first exon and upstream promoter based on this set of quadratic discriminant functions. For example the regions labeled 1 and 2 in figure 2 are predicted by FirstEF to be promoters for the *NPHS1* gene, and region1 is also predicted to be a promoter for *DKFZp564A1164* (although in the reverse orientation). No independent promoter was predicted for

HSPOX1 by FirstEF, however, the similarity in expression patterns between the NPHS1, *DKFZp564A1164* and *HSPOX1* genes led us to hypothesize that these genes may share a single promoter. The potential sharing of a single bi-directional promoter region by three neighboring genes made this region especially interesting to analyze.

Bioluminescent reporter assay

To analyze the functionality of this possible bi-directional promoter as well as other putative promoters in the *NPHS1* gene region, a transient luciferase reporter assay was used. Bioluminescent reporter assays have a wide range of applications including the functional analysis of promoters and enhancers, and it has been demonstrated that these systems provide reliable reproducible results (Parsons, 2000; Sherf, 1996).

The Dual-luciferase reporter system (Promega Corporation) utilizes firefly and Renilla luciferase in a co-reporter system where Renilla is an internal control allowing for normalization of the firefly luciferase data. In this study, the regions predicted to be promoters by FirstEF were placed into vectors that express firefly luciferase when bordered by a functioning promoter and transfected into the appropriate eukaryotic cell lines.

Preliminary expression data were used as a guide in choosing the appropriate cell lines for our transient reporter assay studies. Expression profiles for these genes were obtained from a number of sources including the Genbank's SAGE and EST databases (serial analysis of gene expression, and expressed sequence tag, respectively, http://www.ncbi.nlm.nih.gov/sage), a database of gene expression using microarrays called Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi), and tissue section *in situ* hybridization analysis that was performed at Lawrence Livermore National Laboratory (LLNL). Finally, the most likely candidate cell lines were tested for expression of the genes of interest using RT-PCR (reverse transcription-polymerase chain reaction) and gene specific primers.

Thesis Objective

This thesis' primary objective was to use comparative sequence analysis programs such as PipMaker and VISTA, in addition to the computational program FirstEF, to identify potential promoters and enhancers for three genes in the *NPHS1* region, and to test these regulatory elements in cultured mammalian cell lines using transiently expressed luciferase reporter constructs. Additionally, determining the first exons for *NPHS1*, *HSPOX1* and *DKFZp564A1164,* including potential alternative start sites linked to different promoters was attempted and results sequenced. Overall the aim has been to test the hypothesis that a single bi-directional promoter was being shared by *NPHS1, DKFZp564A1164* and *HSPOX1,* three neighboring genes with similar expression patterns, and to establish the technology and methods for a high throughput assay of promoter and enhancer elements.

MATERIALS AND METHODS

Sequence Comparisons

An 845 kb contig from human chromosome 19 (Genbank accession no. NT_011296) and related regions in mouse (Genbank accession nos. AC087141 and AC020839) were compared using the PipMaker program (http://bio.cse.psu.edu/pipmaker) (Schwartz *et al*, 2000) . In species that diverged 100-300 million years ago, such as human and mice, exons and gene regulatory elements are detectable as similar sequences. These can be visualized on a percent identity plot (PIP), which shows the position in one sequence and degree of similarity between the aligning sequences (Schwartz *et al*, 2000) . In collaboration with M. Zhang (Cold Spring Harbor Laboratory), FirstEF predictions were used to analyze the sequence, and regions predicted to be hypothetical promoters by FirstEF were further analyzed for promoter activity.

Cell Culture

Human and mouse cell lines from American Type Culture Collection (ATCC) were cultured in media and sera recommended by ATCC and containing 100 I.U./ml of penicillin, 100 µg/ml streptomycin and 2 mM of L-glutamine. Growing cultures were housed in a cell culture incubator at 37˚C with 5% CO_2 or as recommended. We preliminarily selected the cell lines based on publicly available SAGE expression data (NCBI) for HSA19 genes, for growth characteristics, for transfection assay performance (based on our own results and published data), and to represent a wide variety of cell types and tissues.

Analysis of cDNA

To determine which cell lines express the genes of interest, RNA was collected from the most likely cell candidates based on expression data obtained on public databases or previous studies, and cDNA was produced via RT-PCR using the RNAqueous kit (Ambion Inc.). Cells were grown as recommended by ATCC until they reached a yield of 1×10^5 to 10^{8} , then the cells were collected and stored in RNAlater (Ambion Inc.) until cDNA was made. Primers were developed that specifically amplified the 3'ends of the cDNA of interest, and standard PCR was performed using Perkin Elmer reagents on an MJ Research thermocycler. Primer sequences are listed in table 1A in the Appendix. If a band was produced of the expected size, then that cell line was considered to express the gene and was used in subsequent transfection assay experiments.

5′ End Transcript Verification

In the case of *HSPOX1* where FirstEF and other methods, such as the presence of CpG islands or GATA and TATA boxes, did not predict a promoter and first exon, 5′ SMART RACE (BD Biosciences Clontech) was performed to verify the position of the first exon. SMART RACE incorporates a switchingm echanism at the 5′ end of an RNA transcript coupled with RACE (rapid amplification of cDNA ends) to isolate the complete 5′ end sequence of a target gene. Additionally 5′ SMART RACE was performed on

NPHS1 as FirstEF predicted 2 first exons for this gene. Often it is the case that the transcription start site is upstream from the start ATG codon in an untranslated initial exon. It was the hoped that 5′ RACE would help to identify any possible untranslated initial exons, and also to establish the sequence of the proximal promoter. After performing 5′ SMART RACE the PCR product was subcloned into a TA vector (Invitrogen Corp,) and sequenced using vector primers [m13(-20) and m13] on an ABI Prism 377 sequencer.

Construct Development

Vector preparation

The pGL3 enhancer or promoter vectors (Promega Corporation) were double digested overnight with the appropriate restriction enzymes (MluI and BglII or KpnI and BglII from New England Biolabs, Inc.) for directional subcloning, then the vector was dephosphorylated to prevent recircularization using alkaline phosphatase from calf intestine (New England Biolabs, Inc.). Following which the vector was purified from an agarose gel using a Qiagen kit and eluted in TE. A test of the vector's re-ligation efficiency was performed by transforming Electromax cells (Gibco Invitrogen Corporation) and growing on an LB/AMP plate overnight. Vectors were considered good if less than 75 colonies grew.

Insert preparation

Primers were designed that flanked the hypothetical promoters and contain restriction sites at the 5′ end complementary to the sites in the vector's multi-cloning site.

Then PCR was performed and a small aliquot run on a gel to determine that the PCR worked. The PCR product was treated with Klenow fragment (New England Biolabs, Inc.) to fill in 3′ recessed ends, and then the PCR product was double digested with the appropriate restriction enzymes and gel purified.

Ligation

The pGL3-Enhancer or -Basic vector and insert were ligated with T4 DNA ligase (New England Biolabs, Inc.) overnight using an excess of insert. Electromax cells were transformed with the ligation product and plated overnight on LB/AMP after out growth for 1 hour in LB only. Colonies were screened via PCR using vector specific primers, and those that contained the insert were grown in LB/AMP overnight and isolated using the Qiagen High Speed Midiprep. An aliquot of the isolated constructs was confirmed by restriction digestion or nested PCR and later sequenced.

Transfection Assays

Dual Luciferase Transfection Assays (Promega Corporation) were performed to determine if the predicted promoters functioned *in vitro*. Bioluminescent reporter assays have been demonstrated to provide reliable reproducible results for the functional analysis of promoters and enhancers (Parsons, 2000; Sherf, 1996). Promoter assays were performed using the pGL3 -Enhancer vector and internal control co-reporter, pRL-TK (Promega Corporation). Promoter and enhancer assays were performed using the pGL3-Basic vector and the same internal control co-reporter.

pGL3-Enhancer Vector

The pGL3-Enhancer vector contains *luc*+ cDNA, which encodes modified firefly luciferase, a multiple cloning region upstream of *luc*+ for insertion of the promoter element, and an SV40 enhancer located downstream of *luc* +. The SV40 enhancer aids in the verification of functional promoter elements by increasing the levels of *luc*+ transcription.

pGL3-Basic Vector

The pGL3-Basic vector contains *luc*+ cDNA, which encodes modified firefly luciferase, and a multiple cloning region upstream of *luc*+ for insertion of the promoter+enhancer element. The pGL3-Basic vector does not contain an SV40 enhancer or promoter in order to determine the presence of a functional enhancer and promoter in the experimental construct.

pRL-TK Vector

The pRL-TK vector is an internal control reporter intended to be used in combination with any experimental reporter vector to co-transfect mammalian cells. The pRL reporter vector contains a cDNA (*Rluc*) encoding Renilla luciferase, which was originally cloned from the marine organism *Renilla reniformis* (sea pansy). The pRL-TK vector also contains the herpes simplex virus thymidine kinase (HSV-TK) promoter to provide low to moderate levels of Renilla luciferase expression in co-transfected mammalian cells.

pGL-Control Vector

The pGL-Control vector contains the SV40 promoter and enhancer sequences, resulting in strong expression of *luc*+ in many mammalian cell types. This is useful in monitoring transfection efficiency in general and is a convenient internal standard for promoter and enhancer activity. The specific transcriptional activity of pGL vectors varies for different cell types and the pGL-Control vector can help determine activity to be expected by a strong promoter.

Transfection

Human cell lines HepG2 and 293, determined to express the gene of interest by analysis of cellular cDNA with gene specific primers, were plated in a 96 well format. One hundred microliters of cells were plated in Opti-MEM (Gibco BRL) at 1×10^4 cells per well in the center 60 wells. The outer wells were filled with 100µl of PBS to prevent drying. Twenty-four hours later the cells were transfected with the vectors via lipofection according to the Fugene6 Transfection Reagent protocol (Roche Molecular Biochemicals). For each well, 5.52×10^{-14} moles of experimental vector (pGL3 plus insert) and 50ng of co-reporter (pRL) were mixed with 0.3 – 1.8 µl of Fugene in 5µl Opti-MEM and added to each well. Moles were chosen as the measuring unit for the experimental constructs to help ensure an equal amount of each construct was delivered

Constructs ranged in size from 5.5 to 8.0 Kb. The plates were then incubated for approximately 24 hours before the Dual Luciferase assay was performed.

Dual Luciferase Assay

The assays were conducted according to Promega's Dual Luciferase® Reporter 1000 Assay system. The media from each well was removed, and 20µl of PLB lysate was added to each well. The plate was then incubated at 37˚C with shaking until the cells were completely lysed. Then 100µl of LARII was added, activating any firefly luciferase generated by the pGL3 construct. The RLU (relative light units) of firefly luciferase was measured using a Packard LumiCount microplate luminometer set to a 5 second read time and a 1 second delay between reads. Then 100µl of Stop&Glo reagent was added to each well and the RLU of *Renilla* luciferase from the pRL-TK co-reporter was measured.

Data Analysis

Each construct was transfected in 3 wells and each well was measured in triplicate. Firefly measurements were averaged for each construct, and Renilla measurements were also averaged per construct in the same manner. The negative control treatment contained pGL-Enhancer or pGL3-Basic construct without insert co-transfected with pRL-TK. The positive control treatment contained pGL-Control and was used as a comparison to maximum expression. The measurements for these were averaged in the same way. The change in fold activity was determined by dividing the sample ratio by the negative control ratio (firefly avg. RLU divided by Renilla avg.

RLU). Constructs that caused an increase in fold activity above the negative controls were considered to contain a working promoter.

rVISTA Analysis

Analysis using rVISTA (http://www-gsd.lbl.gov/vista/) was performed on all the constructs developed and transfected to better understand which transcription factor binding sites may be contained within the constructs and therefore which transcription factors may be acting on the sequences. rVISTA is a computational tool for comparative sequence-based discovery of functional transcription factor binding sites (TFBS) (Loots, 2002).

More specifically, rVISTA enables the high throughput detection of cis-regulatory elements by combining clustering and analysis of conserved interspecies sequence to maximize the identification of functional sites. Initially rVISTA aligns human and mouse sequences using AVID, a global alignment program. Then potential transcription factor binding sites are predicted by Match™ program based on TRANSFAC Professional library 5.3. After finding all the TFBS in each species independently, the sites where core positions correspond in both species are selected as aligned sites. Finally, only the aligned transcription factor binding sites that are found within conserved human-mouse sequence at a level of 80% or more are selected by rVISTA as probable transcription factor binding sites.

RESULTS

Preliminary Expression Data

Expression data was used as a guide in choosing the appropriate cell lines for our transient reporter assay studies. Expression profiles for these genes were obtained from a number of sources including the Genbank's SAGE and EST databases as well as Gene Expression Atlas' microarray database. Additionally, tissue array analysis was also performed at LLNL.

SAGE and EST

The SAGE database uses a technique, which quantifies a "tag" that represents the transcription product of a gene. The number of times a particular tag is observed provides the expression level of the corresponding transcript. The histogram denotes expression level. Using the SAGE histogram as a guide, the strongest expression of *NPHS1* was found in the kidney, brain, mammary gland and testis tissues.
SAGE expression data also showed *HSPOX1* to be expressed in kidney and normal liver tissue.

The EST database showed *NPHS1* expression in the endometrium, adenocarcinoma cell line and Islets of Langerhans; and *HSPOX1* expression in liver, spleen and kidney. Both Genbank's SAGE and EST expression data showed *DKFZp564A1164* to be expressed in brain, germ cells, kidney and lung.

Gene Expression Atlas

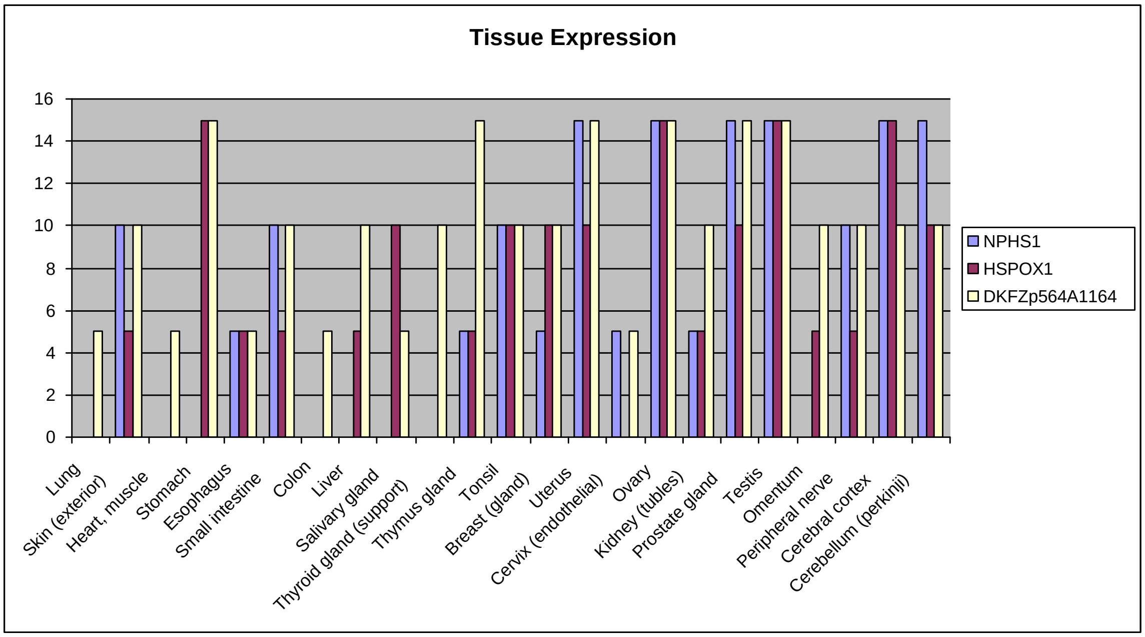
Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi) microarray database only contained data for 2 of the genes we tested (*NPHS1, HSPOX1)*. For *NPHS1* strong positives (3×above median) were noted in the DOHH2, lymphoma and B cell lines, and in the kidney, pituitary and pancreas (10×above median) tissues. For *HSPOX1* positives, all 10×above median, were noted in 3 tissues: kidney, fetal liver and liver. There were no positives 3×above median for *HSPOX1*. (Appendix, Figures1A and 2A)

Tissue Array Results

For the tissue array analysis *NPHS1*, *HSPOX1* and *DKFZp564A1164* genes were hybridized to human tissue array slides by X. Lu and E. Wehri, at LLNL. T7 mRNA probes were made using the mRNA sequence of each gene and ordered from Life Technologies (GibcoBRL). The sequence of each probe is listed in table 2A in the Appendix. The probes were then labeled with dig and hybridized to MaxArray normal human tissue slides (Zymed Laboratories, Inc.) using standard protocols.

The results in graph 1 bellow indicate positive expression in several tissues, ranging in level from 5 to 15 (arbitrary values). For example, all three genes are highly expressed in the testis and ovary and moderately expressed in the kidney tubules. *DKFZp564A1164* and *HSPOX1* were expressed in liver while only *DKFZp564A1164* was expressed in lung, heart and colon. None of the genes were expressed in the spleen or skeletal muscle. Table 1A in the Appendix shows all the tissues tested and the expression result. Figures 3A, 4A and 5A (Appendix) are pictures of the hybridization results on liver tissue. From these results one can see that in the liver *NPHS1* is very weakly expressed, *HSPOX1* is moderately positive and *DKFZp564A1164* is positive and expressing the gene in specific cells of the liver.

The results of the tissue array experiments are unique in that they can show the type of cell within a tissue that is expressing the gene. More often than not a gene is expressed in a specific cell type in the tissue and not the whole tissue. In kidney for example, the expression of*DKFZp564A1164* and *HSPOX1* were only seen in the cells lining the tubules (data not shown). For this reason the data are not always the same as other expression studies where results from a whole tissue or individual cell line are examined.



Graph 1: Comparison of positive tissue hybridization results.

cDNA Analysis and RT-PCR Results

PCR was performed on cDNA made from RT-PCR of the individual cell lines or RT-PCR of poly A$^+$ RNA purchased from BD Biosciences. Commercial cDNA from Clontech was also tested as a control. Samples were run on a 1.2% agarose gel containing 25μg of ethidium bromide at 90V for 30 minutes. Gels were then imaged using the Alpha Imager 2000. Cell lines were considered to be expressing the gene if a band of the expected size was seen on an agarose gel.

All the primers were designed from the 3' end of the cDNA to span an intron so that a size difference could be visualized between genomic and cDNA. *NPHS1* cDNA size was 273bp and genomic DNA was 517bp, *DKFZp564A1164* cDNA size was 391bp and genomic DNA was 3Kb, likewise *HSPOX1* cDNA size was 306bp and genomic was 3Kb. As a positive control primers amplifying β-actin were used, and a PCR reaction lacking any template was used as the negative control.

NPHS1 and *DKFZp564A1164* were found to be expressed in several human cell lines including 293, MDA MB-436, and PANC-1, as seen in figure 4 and table1. Alternatively, expression of *HSPOX1* was only found in 2 of the human cell lines tested, Capan-1 and HepG2. Although *HSPOX1* is expressed in human kidney tissue, there was no indication of expression in the human kidney cell line 293 using this method**.** This may be due to the fact that expression data from RT-PCR of individual cell lines often differs from tissue analysis due to the difficulty of maintaining the tissues differentiated function *in vitro* (Mather & Roberts, 1998).

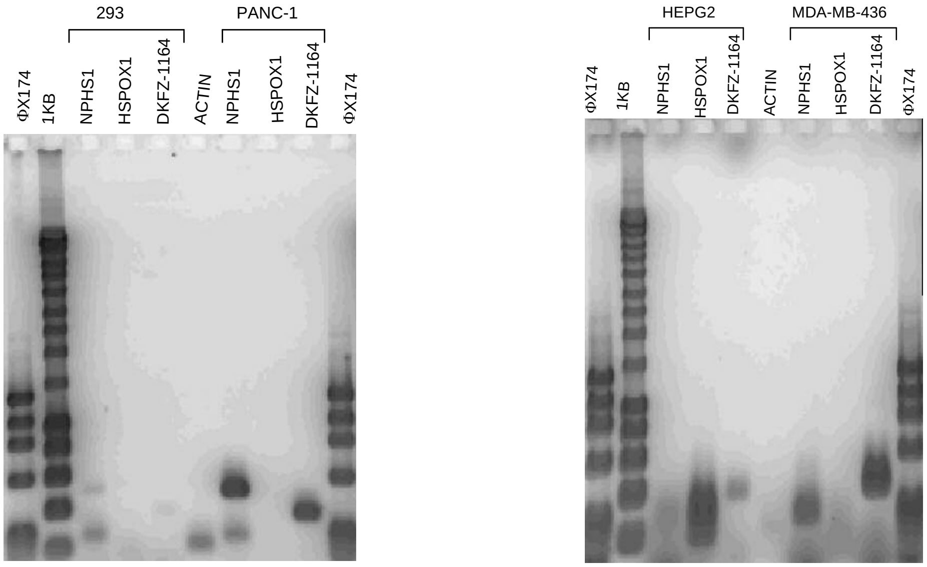


Figure 4: PCR of cDNA from 293, PANC-1, HepG2 and MDA-MB-436 cell lines. Bands representing *NPHS1* and *DKFZp564A1164* expression are seen in 293, PANC-1 and MDA-MB-436. *HSPOX1* and *DKFZp564A1164* expression are seen in HepG2. Actin is a positive control.

Based on this information HepG2 and 293 cell lines were chosen to be used in the transient transfection luciferase assays because of their clear unambiguous results and *HSPOX1*and *NPHS1* which may be sharing a bi-directional promoter, are differentially expressed in these cell lines. LNCaP.FGC, which did not show expression of any of these genes was used to test luciferase assay results in a non-expressing cell line.

Table 1: RT-PCR Expression Data

Cell Line or Tissue (human)	Tissue	Gene	cDNA present
293	kidney	1164**	yes
293	kidney	NPHS1	yes
293	kidney	HSPOX1	no
Capan-1	pancreas	All*	yes
Hela S3	cervix	HSPOX1	no
Hela S3	cervix	1164	faint band
Hela S3	cervix	NPHS1	yes
HepG2	liver	1164**	yes
HepG2	liver	HSPOX1	yes
HepG2	liver	NPHS1	no
IMR-32	neuroblast	All*	no
Jurkart	leukemia, T-cell	All*	no
k562	leukemia	All*	no
LNCaP.FGC	prostate	All*	no
MDA-MB-436	breast	1164**	yes
MDA-MB-436	breast	HSPOX1	no
MDA-MB-436	breast	NPHS1	yes
MDA-MB-453	mammary	All*	no
PANC-1	pancreas	1164**	yes
PANC-1	pancreas	HSPOX1	no
PANC-1	pancreas	NPHS1	yes
commercial RNA	kidney	HSPOX1	yes
commercial RNA	kidney	1164**	no
commercial RNA	kidney	NPHS1	yes
commercial cDNA	brain/testis	NPHS1	yes
commercial cDNA	brain/testis	1164**	yes
commercial cDNA	brain/testis	HSPOX1	no

*DKFZp564A1164, NPHS1, HSPOX1
**DKFZp564A1164

5′ End Transcript Verification

Five prime RACE (SMART RACE, BD Biosciences Clontech) was performed to verify the position of the first exon for both *HSPOX1* and *NPHS1*. Often it is the case that the transcription start site is upstream from the start ATG codon in an untranslated initial exon. The 5′ RACE experiment served to identify a possible untranslated initial exon, and therefore also to establish the position of the proximal promoter. After performing 5′RACE the PCR product was subcloned into a TA vector (Invitrogen Corp,) and sequenced using vector primers on an ABI Prism 377 sequencer.

As starting materials commercial liver and kidney poly A$^+$ RNA from BD Biosciences Clontech were used. These RNAs were initially tested for the presence of the *HSPOX1* and *NPHS1* cDNAs using the same primers designed to test the cell line RNA in the cDNA analysis method above.

The initial results from the 5′ RACE were inconclusive. After several separate SMART RACE experiments, the 5′ regions of both *HSPOX1* and *NPHS1* have still not been identified. Not only were no new untranslated first exons identified, but also the currently accepted 5′ end of theses genes could not be verified using this method. The positive control provided with the kit was used in conjunction with these experiments and did produce the expected results.

In performing the 5′ RACE experiments on the *HSPOX1* gene it was noted that the gene’s first and second exons matched to multiple sites in the genome using NCBI BLAST, and when aligning the human and mouse mRNA sequences, it was found they do not form a consensus sequence alignment until base pair 298 in human which is equivalent to amino acid 77. Even when choosing unique primers from the consensus region, the 5′ end of the gene was not found using SMART RACE. RACE products were generated but sequence did not correspond to any sequence from this genomic region.

Although the *NPHS1* gene is well characterized, and the mouse consensus region matches well, the 5′ end of the gene was not established using SMART RACE. The sequence of *NPHS1* RACE products matched ring finger/*DORFIN*, crystalin/*CRYL1*, glutathion/*GSTA2* and ribonuclease/*PARN*, indicating that false priming was generating artifacts from abundant RNAs in the sample.

Primer design was of critical importance in these experiments. The 30 bp primers designed for these experiments had to match the gene of interest exclusively; if any part also matched a different area of the genome one risked amplifying both regions. Careful screening of not just the whole gene specific primer, but small segments of the primer was therefore necessary. BLAST searches revealed that the exons of these genes (*HSPOX1* and *NPHS1)* are littered with small sequence segments of 10 to 20 bp in length that match other regions of the genome, making it difficult to find 30 bp gene specific primers for the SMART RACE experiments (primer sequences: Table 5A, Appendix). These repeat sequences most likely explain the failure of RACE to generate *NPHS1* and *HSPOX1* specific transcripts.

Under these circumstances the published 5' ends are probably the true ends of these transcripts, at least in the cell types tested. Since certain promoters may operate only in specific tissue types, it is possible that exhaustive RACE in many tissues would have eventually yielded additional 5' sequences. However, such a search was beyond the scope of this study.

Transfection Assays

Dual Luciferase Transfection Assays (Promega Corporation) were performed to determine if the FirstEF-predicted promoters functioned as promoters *in vitro*. Bioluminescent reporter assays have been demonstrated to provide reliable reproducible results for the functional analysis of promoters and enhancers (Parsons, 2000; Sherf, 1996). Promoter assays were performed using the pGL3 -Enhancer vector and internal control co-reporter, pRL-TK (Promega Corporation). Promoter and enhancer assays were performed using the pGL3- Basic vector and the same internal control co-reporter.

Preliminary Transfection Data: *XRCC1*

In order to determine the effectiveness of the Promega's Dual Luciferase Assay the *XRCC1* gene was shotgun subcloned into the pGL-Enhancer vector. By aligning the baboon *XRCC1* promoter sequence (Genbank accession no. AF019114), which had been previously cloned and characterized by Zhou *et. al,* with human (Genbank accession no. L34079) and mouse (Genbank accession no. L34078) using mVISTA we were able visualize the human promoter region (Figure 5) (Zhou & Walter, 1998).

Then webcutter (http://www.firstmarket.com/cutter/cut2.html) was used to determine which restriction enzyme would be best to use, and the human clone (Genbank accession no. L34079) was digested with SacI. The SacI digest resulted in seven fragments, all of which were shotgun subcloned into the pGL3-Enhancer vector. Colonies were isolated that had 3.7kb, 3.8kb, 3.9kb and 7.9kb inserts, and these were tested using Promega's Dual Luciferase Assay. Promega's pGL3-Control, which contains a SV40 constitutive promoter, was used as a positive control, and an empty pGL3-Enhancer was used as the negative control (Figure 6). The fragments were also sequenced and positional verified.

The results showed that only the vector containing the promoter worked. The other shotgun subcloned sequences had values similar to the negative control demonstrating that the assay does not typically generate false positives.

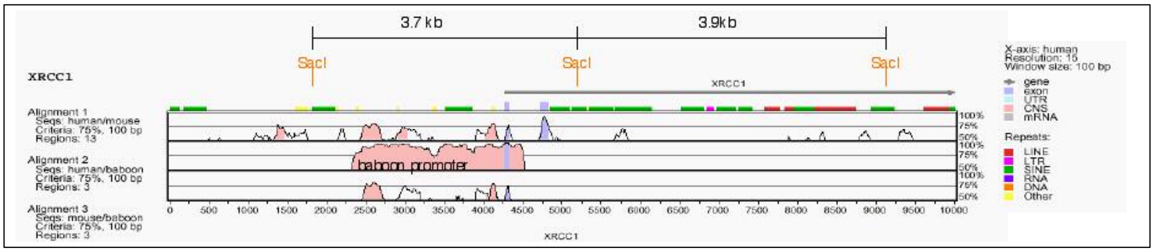


Figure 5: Alignment of human, baboon and mouse sequence using mVISTA (Genbank accession nos. L34079, AF019114, L34078 respectively).

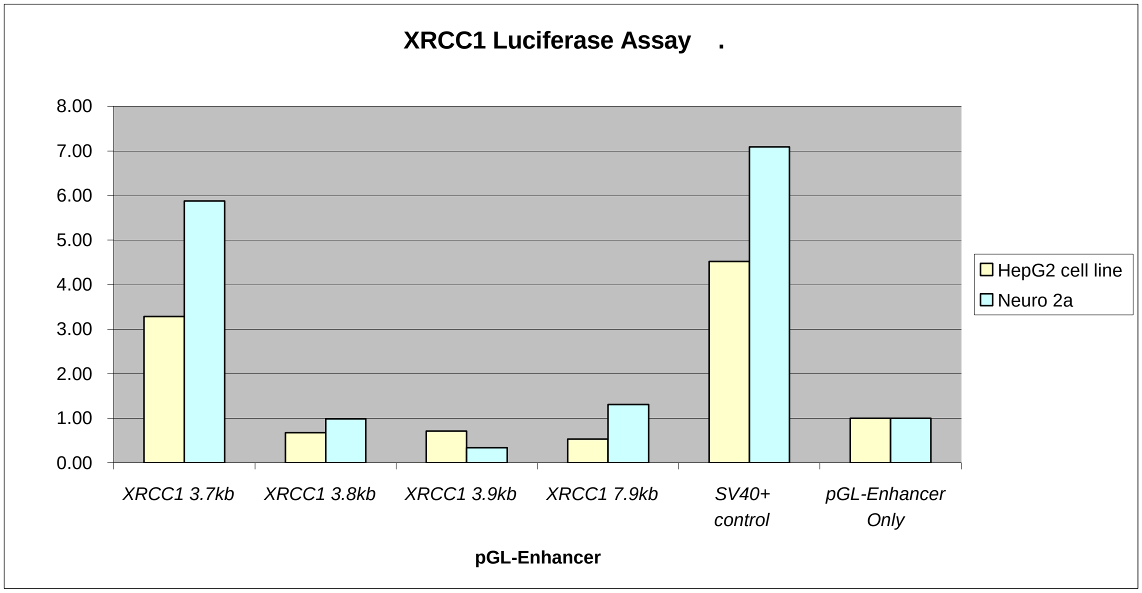


Figure 6: Luciferase Assay of *XRCC1* shotgun subcloned fragments.

Construct Design: *NPHS1*, *HSPOX1* and *DKFZp564A1164*

The pGL3 -Enhancer or -Basic vectors (Promega Corporation) were double digested with restriction enzymes Kpn I/Bgl II or Mlu I/Bgl II (New England Biolabs, Inc.) for directional subcloning and ligated with inserts that were double digested in the same manner. The restriction enzymes were chosen based upon a screen of each insert to determine which restriction enzyme sites they did not contain (Webcutter 2.0, copyright 1997 Max Heiman, http://www.firstmarket.com/cutter/cut2.html). Figure 7 is an example of insert design and figure 8 shows the region each vector was designed from and names each were given. Table 2 gives additional information about each construct including size and region of cosmid R33502 (Genbank accession no. AC002133) they were cloned from. The primers for each insert were designed with a BglII, KpnI or MluI site added to the 5′ end according to recommendations in New England Biolabs technical literature. Primer sequences are all listed in the appendix table 4A.

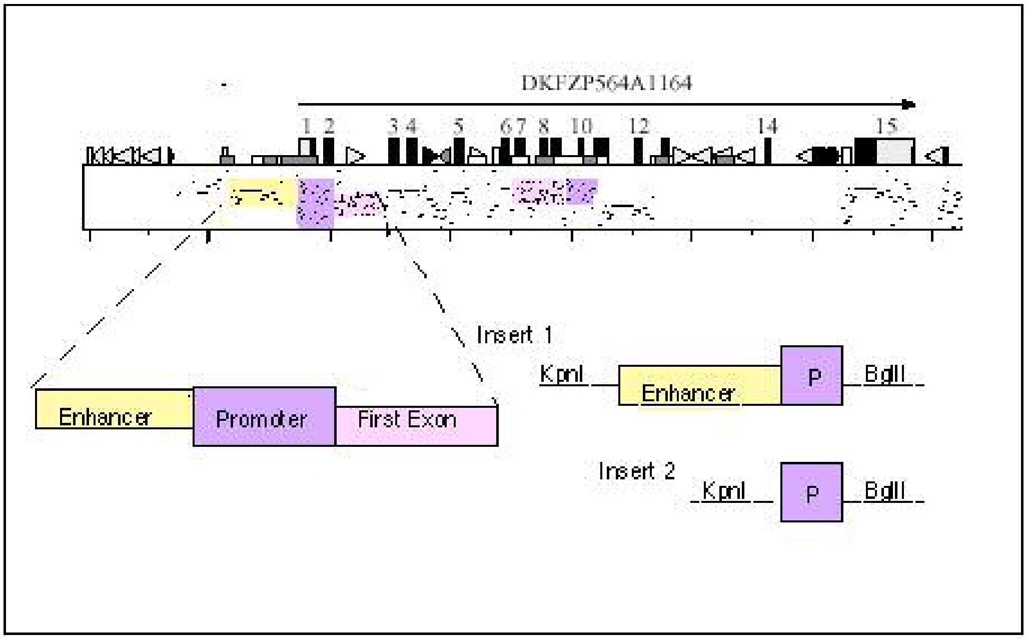


Figure 7: Promoter and enhancer inserts for directional subcloning into pGL3 vectors.

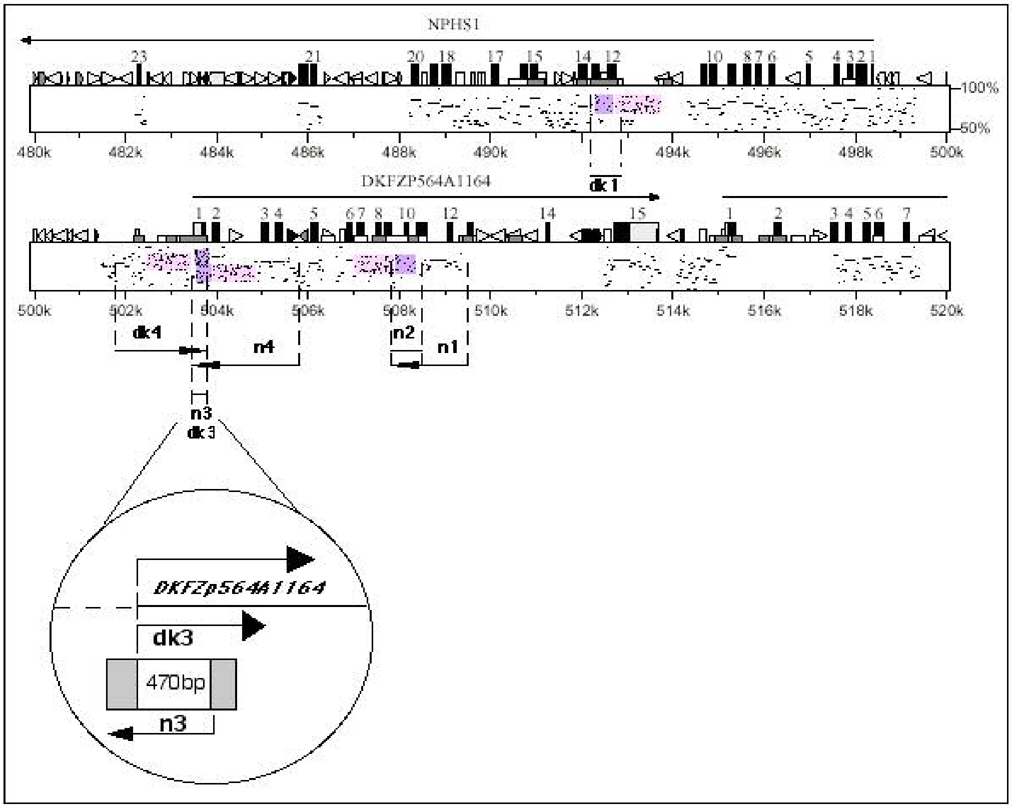


Figure 8: Name of each insert and region it was developed from along the pip plot.

Table 2: Summary of constructs

Name	**Insert size**	**Type**	**Vector**	**Digest**	**Region in cosmid** R33502 (AC002133)
n1-Basic	1.2kb	promoter+enhancer	pGL Basic	BglII/KpnI	37619-36392
n2-Enhancer	570bp	promoter	pGL Enhancer	BglII/KpnI	36964-36392
n3-Enhancer	619bp	promoter	pGL Enhancer	BglII/MluI	32729-32099
n4-Basic	3.2kb	promoter+enhancer	pGL Basic	BglII/MluI	35319-32099
n2r-Enhancer	570bp	promoter reversed	pGL Enhancer	BglII/KpnI	36392-36964
n3r-Enhancer	619bp	promoter reversed	pGL Enhancer	BglII/MluI	32099-32729
dk1-Enhancer	572bp	promoter	pGL Enhancer	BglII/KpnI	20950-21522
dk3-Enhancer	570bp	promoter	pGL Enhancer	BglII/MluI	32232-32802
dk4-Basic	1.6kb	promoter+enhancer	pGL Basic	BglII/MluI	31220-32802
n2-Basic	570bp	promoter	pGL Basic	BglII/KpnI	36964-36392
n3-Basic	619bp	promoter	pGL Basic	BglII/MluI	32729-32099
n2r-Basic	570bp	promoter reversed	pGL Basic	BglII/KpnI	36392-36964
n3r-Basic	619bp	promoter	pGL Basic	BglII/MluI	32099-32729
dk1-Basic	572bp	promoter	pGL Basic	BglII/KpnI	20950-21522
dk3-Basic	570bp	promoter	pGL Basic	BglII/MluI	32232-32802
dk1r-Enhancer	572bp	promoter reversed	pGL Enhancer	BglII/KpnI	21522-20950
dk1r-Basic	572bp	promoter reversed	pGL Basic	BglII/KpnI	21522-20950
dk3r-Enhancer	570bp	promoter reversed	pGL Enhancer	BglII/MluI	32802-32232
dk3r-Basic	570bp	promoter reversed	pGL Basic	BglII/MluI	32802-32232

Human cell lines 293 and HepG2 were determined to express the genes of interest by analysis of cellular cDNA with gene specific primers, and were consequently plated in 96 well format for the transient transfection luciferase assays. LNCaP.FGC, which did not show expression of any of these genes, was used to test luciferase assay results in a non-expressing cell line.

Constructs n1 and n2

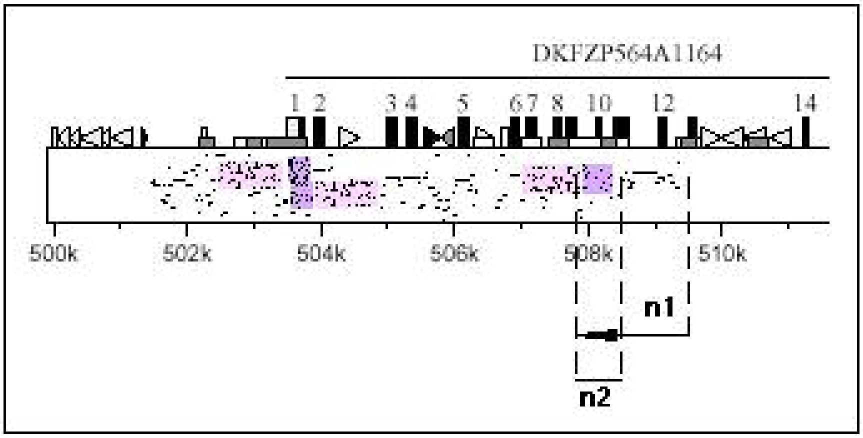


Figure 9: Constructs n1 and n2.

Construct n2 is predicted by FirstEF to be a potential upstream promoter for the *NPHS1* gene. This sequence is positioned unusually for a promoter in that it lies within the *DKFZp564A1164* transcription unit (Figure 9). The n1 construct includes the n2 promoter region plus flanking 630bp of upstream consensus sequence that was considered a possible enhancer region. The n2 promoter was directionally subcloned into the pGL-Enhancer and -Basic vectors to test expression in a construct containing and lacking the SV40 enhancer, respectively. The n1 region was subcloned into the pGL-Basic vector only. The n2r promoter construct is identical to n2 except subcloned in the reverse orientation into pGL-Enhancer and -Basic vectors; this construct was designed as a possible control for n2. Intact pGL-Enhancer and -Basic vectors, which lacked any insert, were used as negative controls, and the pGL-Control vector which contains an SV40 promoter and enhancer was used as an example of a strong positive.

Surprisingly, the results of the transfection assay indicated that both n2 and n2r have strong promoter activity in the pGL-Enhancer vector transfected into the HepG2 cell line. The same constructs also show promote activity in the 293 cell line similar to the positive control, and no activity in LNCap (Figure 10). The strong positive n2 sequence indicates that it may be an alternative upstream promoter for the NPHS1 gene as predicted by FirstEF, and the fact the n2r acts as a strong promoter in both cell lines indicates that it is a bi-directional promoter. These data suggest that the n2r sequence may also function as a downstream alternative promoter for the *DKFZp564A1164* gene. The n1-Basic construct displayed a reduction of promoter activity compared to n2-Basic suggesting there may be a silencer in this region causing repression of expression in the cell lines used for this study. Often silencers causing repression of expression are found in the 5′ upstream region of genes (Kemp *et al.*, 2002; Kraner *et al.*, 1992).

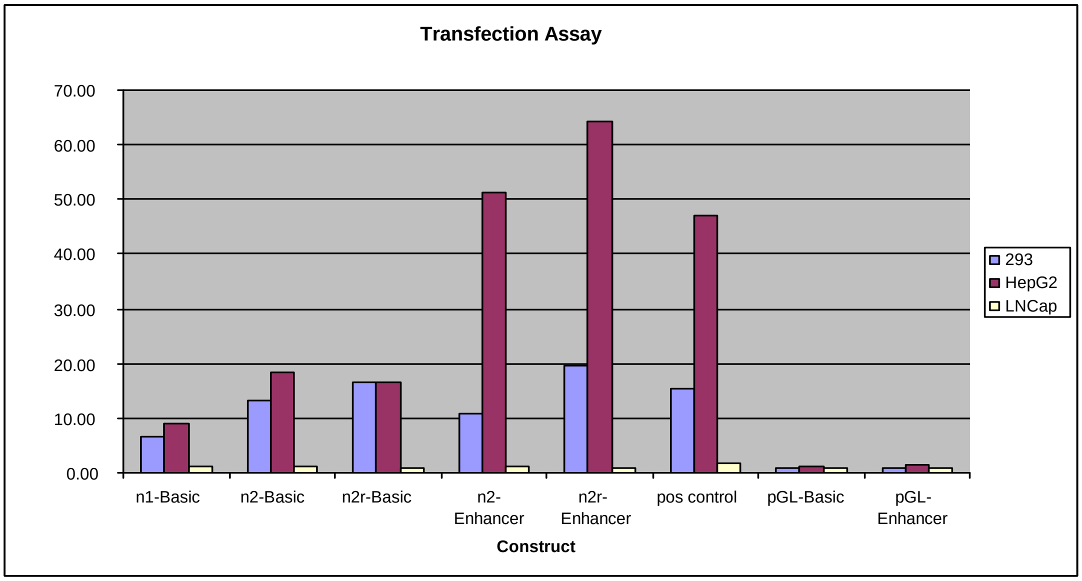


Figure 10: Fold change in relative light units (RLU) of n1 and n2 constructs transfected into 293, HepG2 and LNCap cell lines.

Constructs n3 and n4

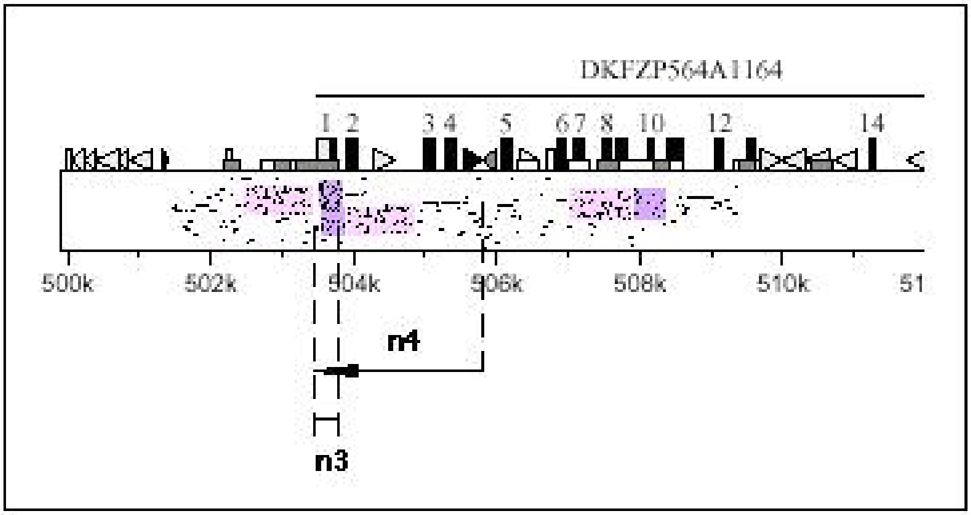


Figure 11: Constructs n3 and n4.

Construct n3 is also predicted by FirstEF to be an alternative potential upstream promoter for the *NPHS1* gene. A growing body of data suggests that many genes use alternate promoters in different tissues(Asnagli *et al.*, 2002). The n4 construct includes the n3 promoter region plus flanking 2581bp of upstream consensus sequence that was considered a possible enhancer region for this promoter (Figure 11). The n3 promoter was directionally subcloned into the pGL-Enhancer and -Basic vectors to test expression in a construct containing and lacking the SV40 enhancer. The n4 region was subcloned into the pGL-Basic vector only. The n3r promoter is identical to n3 except subcloned in the reverse orientation into pGL-Enhancer and -Basic vectors. Intact pGL-Enhancer and -Basic vectors, which lacked any insert, were used as negative controls, and the pGL-Control vector which contains an SV40 promoter and enhancer was used as an example of a strong positive.

The n3 promoter region was predicted by FirstEF to be a potential bi-directional promoter for *NPHS1* and *DKFZp564A1164* and does show higher levels of expression in the reverse orientation (n3r-Basic and n3r-Enhancer) when compared to the negative controls. However, the results of the transfection assay indicate that n3r is a weak promoter in comparison to the positive control (Figure 12). When the scale is decreased in the graph so that differences in promoter activity can be visualized for the test regions, a 4 (n3r-Basic) and 6 (n3r-Enhancer) fold increase in expression is clearly visible (Figure 12 and 13). It should be noted that the positive control used in these experiments was supplied by Promega and contains a very strong SV40 promoter and enhancer, and most human promoters will not be as strong or stronger than the positive control. Expression of the forward *NPHS1* constructs, n3-Basic and –Enhancer were barely 1 fold greater than the negative controls, indicating that n3 is probably not a promoter for the *NPHS1* gene. The n4-Basic construct reduced promoter activity to that seen in the negative controls suggesting there may be a silencer in this region completely shutting off expression.

The difference in expression between cell lines should be noted as well. While the previous constructs always had higher expression in the HepG2 cell line, the n3r promoter shows deferential expression depending on the vector. Expression n3r in the Basic vector was higher in the 293 cell line while expression in the Enhancer vector was higher in the HepG2 cell line. This may just be an artifact of the low expression levels, or an instance of enhancer competition. A study by G.I.R. Adam *et al.* showed that the SV40 enhancer used in many plasmids for transient transfection assays can be a strong competitor for positive and negative regulatory factors in a cell-type-specific manner (Adam *et al.*, 1996). Although the SV40 enhancer clearly performs well in most of the cell types we have examined, this factor may help explain the differences in luciferase levels seen in some cells.

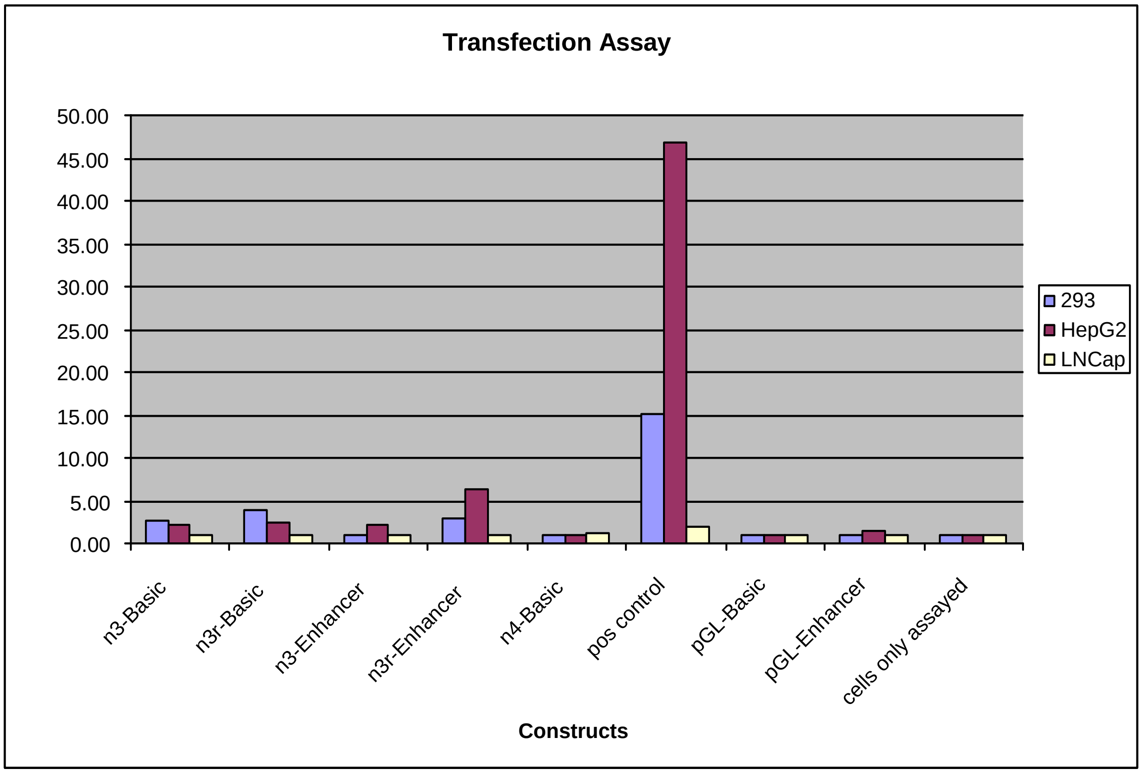


Figure 12: Fold change in relative light units (RLU) of n3 and n4 constructs transfected into 293, HepG2 and LNCap cell lines.

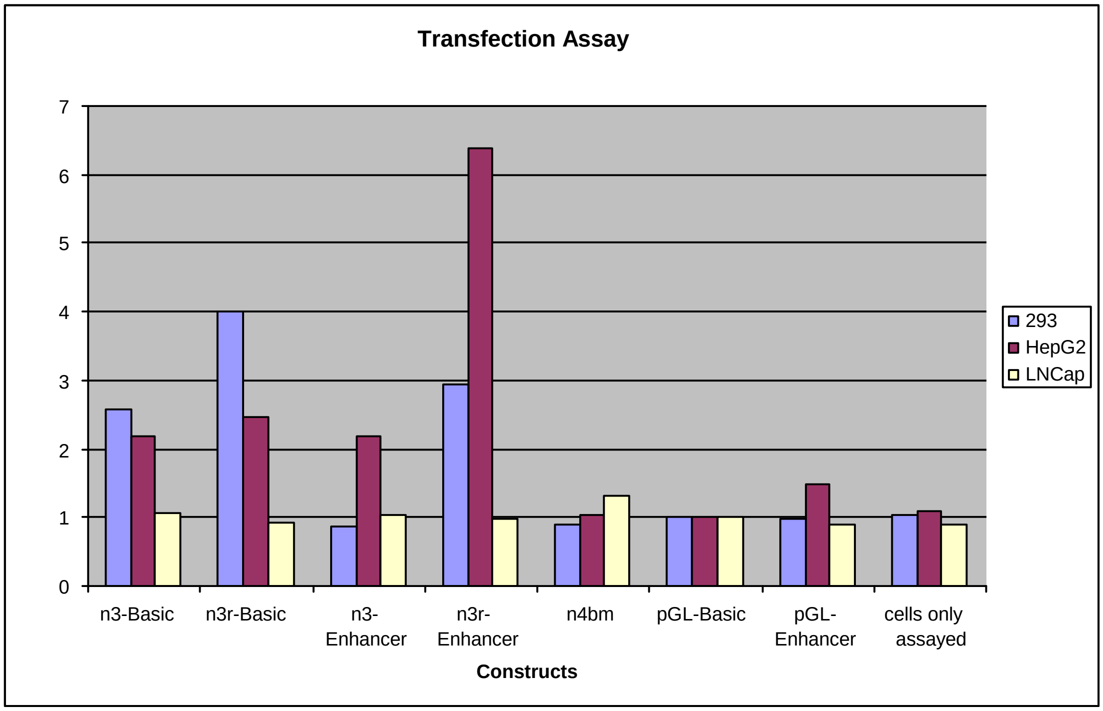


Figure 13: Fold change in relative light units (RLU) of n3 and n4 constructs transfected into 293, HepG2 and LNCap cell lines in comparison to negative controls (positive control removed).

Constructs dk1, dk3 and dk4

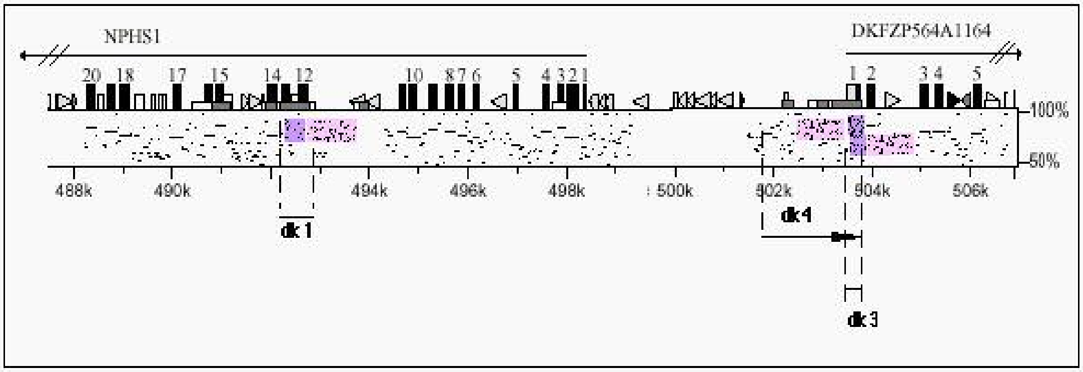


Figure 14: Constructs dk1, dk3 and dk4.

The dk1 and dk3 promoters were predicted by FirstEF to be potential upstream promoters for *DKFZp564A1164* (Figure 14). The dk4 construct includes the dk3 promoter region plus 1kb of flanking upstream consensus sequence that was considered a possible enhancer region. The enhancer region (dk2) flanking dk1 was not subcloned due to difficulties in PCR of this GC-rich region.

Forward promoter constructs dk1-Enhancer and dk3-Basic express luciferase at more than 25 times that of the negative controls in the HepG2 cell line (Figure 15). In the 293 cell line dk1-Basic expresses the highest level of luciferase at almost 10-fold relative light units (RLU). Again the promoter/enhancer construct, dk4-Basic, shows a reduction in luciferase activity compared with the promoter only constructs, suggesting a silencer may be present.

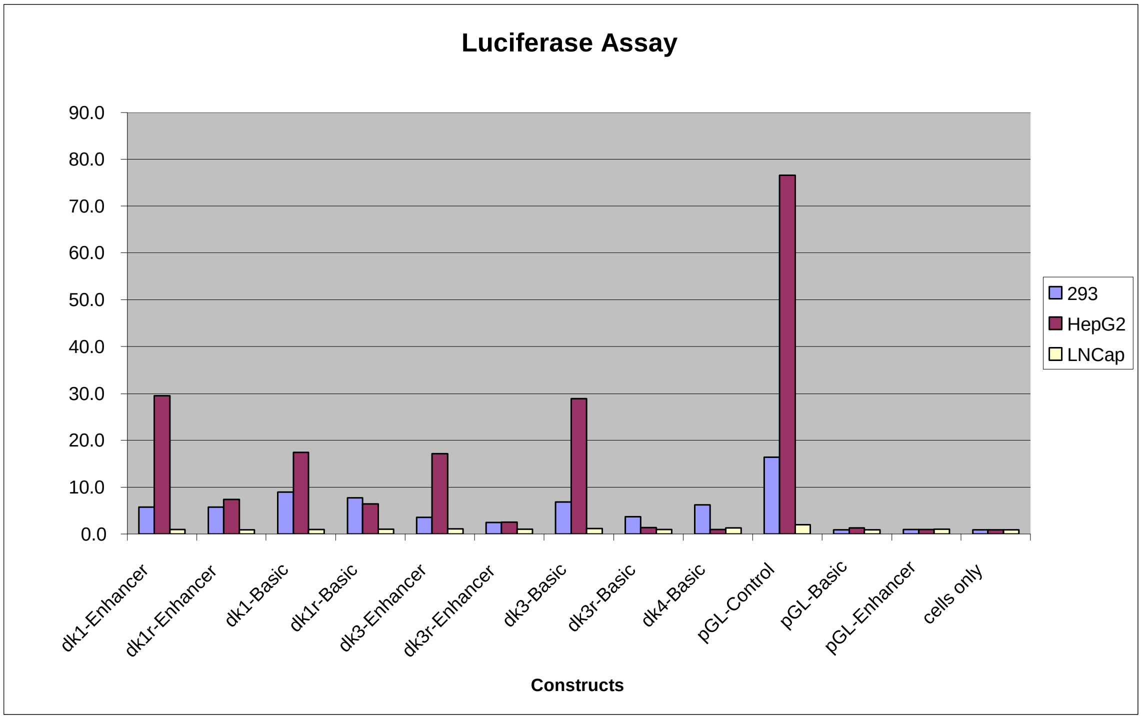


Figure 15: Fold change in relative light units (RLU) of dk1, dk3 and dk4 constructs transfected into 293, HepG2 and LNCap cell lines in comparison to controls.

Constructs dk3 vs. n3

The dk3 and n3 predicted promoters overlap by approximately 470bp and each extends beyond this core region by about 100 bp. The orientation of n3 and dk3 are opposite to each other, whereas n3 is in the same orientation as dk3r, and dk3 is in the same orientation as n3r (Figure 16).

Expression levels were highest for the dk3 constructs (5 to 15-fold increases). In spite of the overlap region, the n3r constructs only showed a 4-fold increase in expression (Figure 17). The dk3r and n3 constructs had the lowest expression levels, similar to the negative controls.

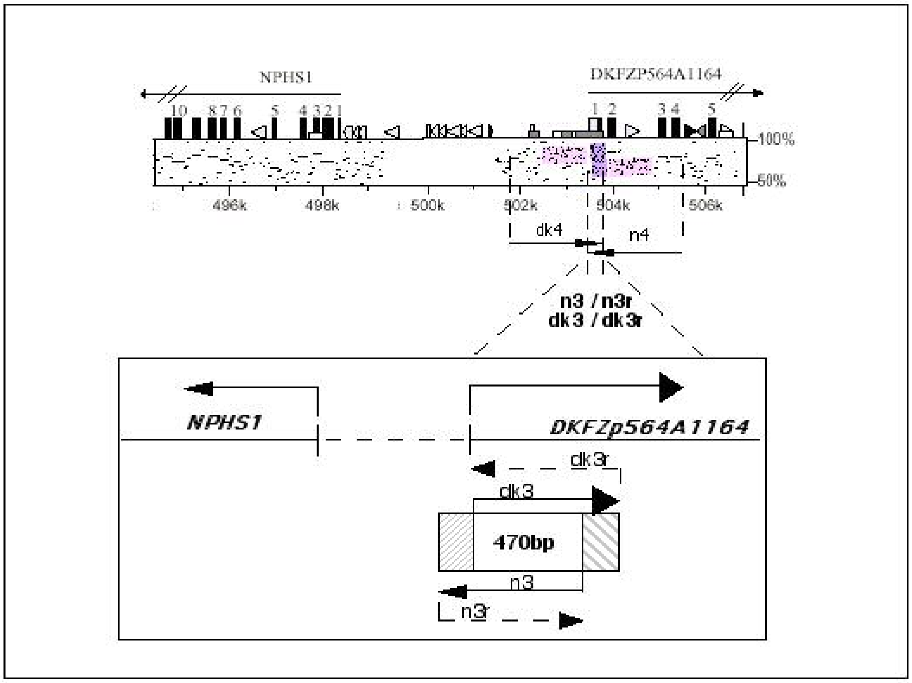


Figure 16: Overlap region of the n3 and dk3 promoters.

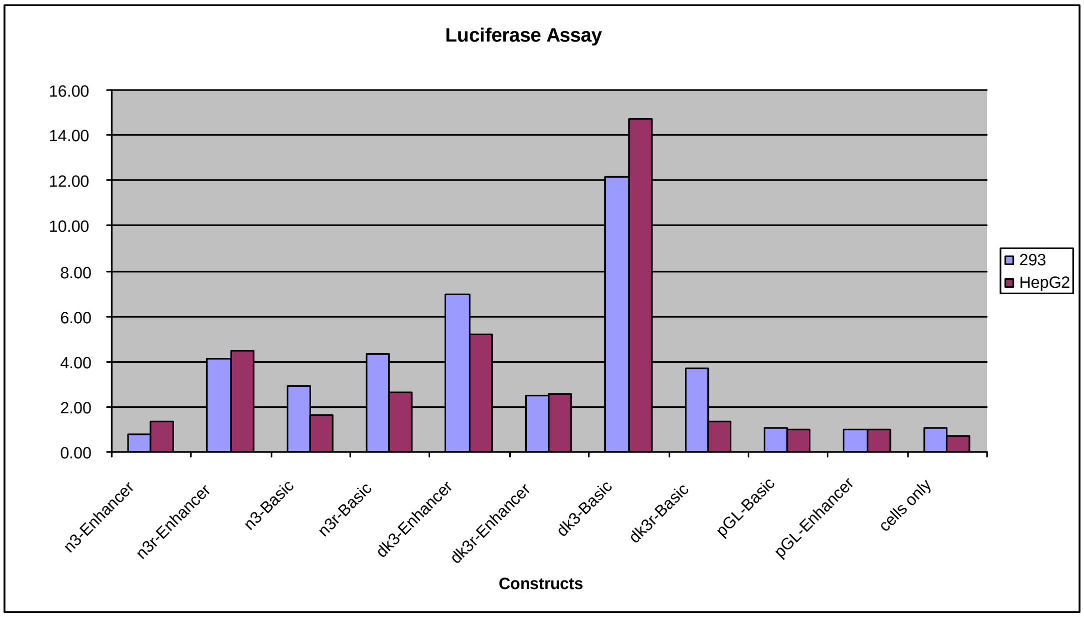


Figure 17: Fold change in relative light units (RLU) of dk3 and n3 constructs transfected into 293 and HepG2 cell lines in comparison to negative controls.

In order to clarify the expression patterns seen in this region, larger and smaller constructs were designed in the 470 bp overlap region (Figure 18). In the luciferase assay the highest level of activity was seen in the new larger dk3-Basic construct indicating the extra region contains a powerful enhancer driving this promoter (Figure 19). The larger dk3-Enhancer construct did not, however, express luciferase at higher levels than the original dk3-Enhancer. Since the pGL-Enhancer vector contains an SV40 enhancer, it may be competing for regulatory factors, preventing them from binding to the insert DNA (Adam *et al*, 1996) .

The smaller constructs and the larger n3 constructs, as well as the original n3, n3r and dk3r constructs all showed low levels of luciferase expression similar to the negative controls. These data seem to indicate that the working promoter is within dk3 forward construct and only operates in one direction i.e. is not bi-directional as predicted by FirstEF. Additionally, at least two strong enhancers are located 5′ of this promoter as evidenced by the high luciferase expression in the dk3-Basic and large dk3-Basic constructs. The strong putative liver and kidney enhancers in this region deserve further study including the possibility that the SV40 enhancer may be competing for the same transcription factors.

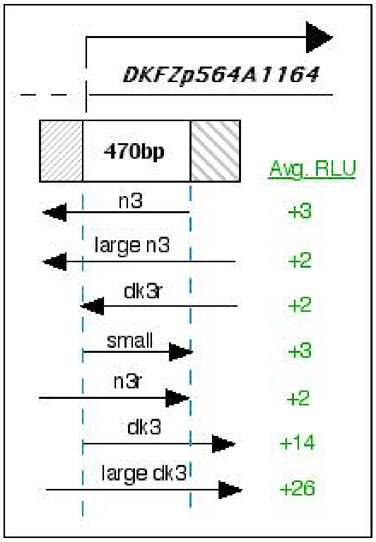


Figure 18: New large and small constructs, their orientation and average RLU.

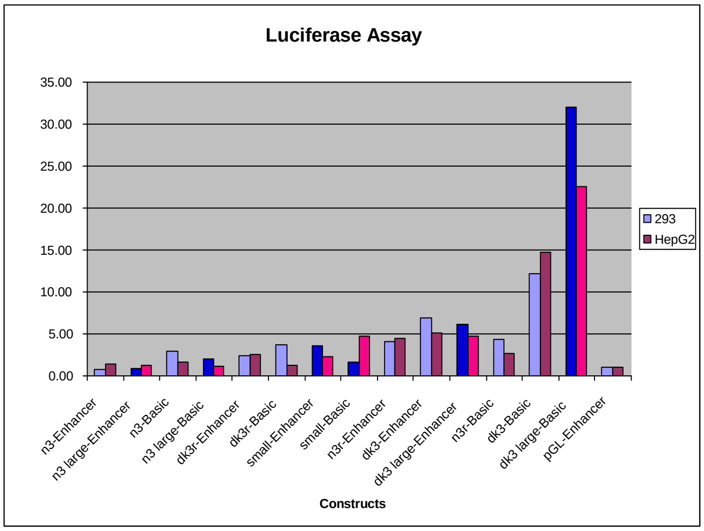


Figure 19: The bight blue and pink bars show the expression patterns of the new large and small constructs.

rVISTA Analysis

rVISTA analysis, which detects transcription factor binding sites (TFBS) by clustering and analysis of conserved interspecies sequence, was performed on construct sequences from n1, n2, n3, n4 and dk3 and dk4 using a core similarity of 0.85 and matrix similarity of 0.9, slightly higher than the default parameters. Construct dk1's similarity standards were left at default, 0.75 and 0.8, respectively. All conserved or aligned TFBS that were found in the sequences are listed in table 6A in the appendix.

The results found 11 conserved TFBS in n1 (enhancer region only) including 1 GATA site and 5 CAP sites. All 11 TFBS are found within a 35 bp region immediately 5′ of the promoter. Promoter n2 contained 12 aligned TFBS including 8 CAP sites and 1 each of CETS1P54, ZIC3, CDXA and MZF1.

Promoter n3 contained 9 conserved TFBS located at the 3′ end, and n4 (promoter + enhancer construct) contained 10 conserved TFBS found in clusters throughout the enhancer region. It has been shown that when multiple cis DNA elements are clustered in a region they may work cooperatively to regulate expression (Belsham & Mellon, 2000; Liu *et al.*, 2003). Some of the transcription factor binding sites found in this region included 22 CAP sites, 7 STAT sites, 8 PAX2 sites, 2 GATA sites and 2 YY1 sites.

In the dk1 promoter, only 2 conserved TFBS were found, CAP and ZP1, and both were located in the 3′ end of the promoter. In the dk3 promoter, 4 transcription factor binding sites were found at the 5′ end including 2 PAX2 sites. Recalling that dk3 and n3 overlap by 470bp, they also share 4 TFBS, and an additional 5 sites are found in the n3 region of this promoter.   When the larger promoter construct incorporating all of n3 and dk3 was assayed the results showed very strong expression in the dk3 orientation in the pGL-Basic vector only, suggesting that the extra TFBS found in the n3 region may actually be enhancers for the dk3 promoter. Twenty TFBS were found throughout in the enhancer region of dk4 including 2 PAX2 sites, 4 STAT, 2 GATA and 6 CAP sites. The high concentration of conserved TFBS, especially the clustering of multiple copies of some TFB sites, are consistent with the predicted enhancer role for this *DKFZp564A1164* region.

DISCUSSION

The results confirm that comparative sequence analysis between divergent species such human and mouse is a powerful tool for identifying regulatory elements in non-coding conserved sequence (Loots *et al*, 2000) . In this study we used the wealth of conserved sequence data for HSA19 and mouse to locate putative promoter elements, and explored the use of comparative sequence analysis programs such as PipMaker or VISTA and the computational promoter finding program, FirstEF, to assist in locating potential promoters. This is the first study designed to test the FirstEF predictions, and the results show that 3 out of 4 predicted promoters were functional in the luciferase assay (Figure 20). However, much larger numbers of FirstEF predictions need to be assayed to assess this method.

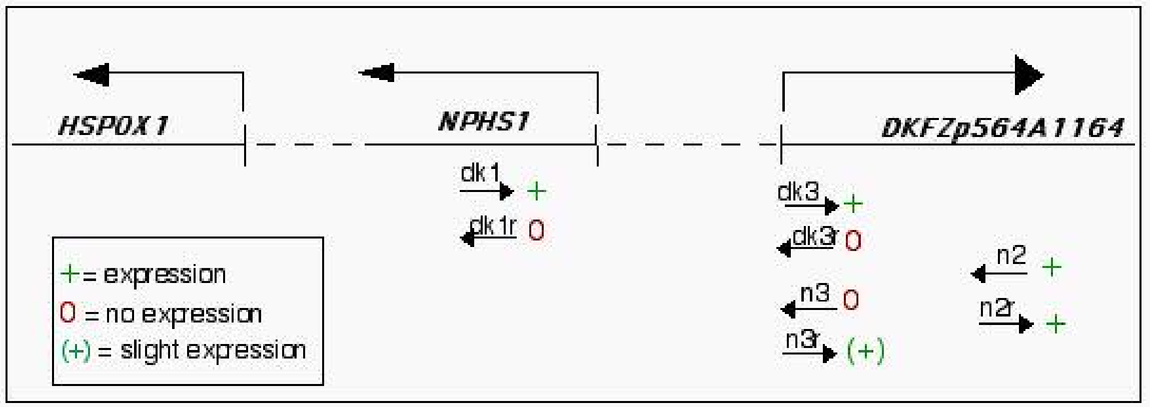


Figure 20: Three of four FirstEF predicted promoters showed expression in the luciferase assay. One predicted promoter (n2/n2r) was found to have expression in both orientations, although it was not predicted by FirstEF to be bi-directional.

The results also demonstrate that testing potential regulatory elements in transiently expressed luciferase reporter constructs transfected into cultured mammalian cell lines is a reliable method, and becomes a high throughput method when performed in a 96 well format. In this study the choice of cell line was found to be of critical importance to assay results. For example in general the HepG2 cell line produced higher luciferase values when transfected with these particular promoters. However, in 3 constructs 293 showed higher values: n3Basic, n3rBasic and dk4Basic. The LNCap cell line, on the other hand, was a poor reporter all together. Many genes use alternative start sites and promoters in different tissues, so promoters should be tested in at least two different cell lines that are based on the results of preliminary expression data (Asnagli *et al.*, 2002).

The results did not, however, show that *NPHS1, HSPOX1* and *DKFZp564A1164* share a single bi-directional promoter (n3/dk3). The n3 construct does not have promoter activity in the cell lines we tested and, therefore, is probably not a promoter for *NPHS1* and *HSPOX1*. However, n2 unexpectedly turned out to be a bi-directional promoter.

The n2 construct is an excellent example of the importance of testing all hypothetical promoters in both orientations. Although, n2 was predicted by FirstEF to be a promoter only in one direction for the *NPHS1* gene it expressed high levels of luciferase activity in both orientations indicating it is a strong bi-directional promoter.

Sequence length was also shown to be of importance in this study because although n3 and dk3 share 470 bp of sequence with each other it was the 70 to 100 base pairs that they did not share that was found to enhance or reduce expression. The rVISTA data showed that several potential transcription factor binding sites (TFBS) exist on the periphery of this core promoter region. When a larger construct was designed it was revealed that the dk3 promoter was further enhanced while the n3 direction remained the same suggesting that the extra base pairs contain TFBS that act as enhancers on the dk3 promoter.

Throughout this project the experiments were formatted to establish the technology and methods for a high throughput assay of promoter and enhancer elements. Using FirstEF as a guide, putative promoters can be quickly assayed for activity. In our study 3 of 4 promoters assayed showed considerable increases in luciferase activity over negative controls denoting a working promoter (Figure 20). Although, construct n3 only showed a slight increase in luciferase activity over the negative controls, dk3, which overlaps the same region, but was cloned in the opposite orientation, showed very strong luciferase activity. Not all mammalian promoters are going to be as strong as positive controls and we should expect to see a high degree of variability in expression.

The consensus sequence upstream of each predicted promoter was tested for enhancer activity in the pGL3-Basic vector, and all 3 of these “enhancer” constructs had reduced activity relative to the shorter promoter sequences. As previous studies have shown the 5′ region of a promoter can contain silencer sites causing transcriptional repression (Kemp *et al.*, 2002; Kraner *et al.*, 1992). When these regions were examined by rVISTA a number of transcription factor binding sites were found, some of which are known to be repressors for certain genes or in specific tissues. For example YY1, PAX2 and CIZ binding sites were found in one or more of the enhancer regions and all have been shown to reduce expression in previous studies (Havik *et al.*, 1999; Kim *et al.*, 2003; Shen *et al.*, 2002).

The enhancer constructs were more difficult to PCR and subclone due to their larger size and high GC content making the generation of these regions more time consuming. For this reason high throughput assays of putative enhancer/silencer regions may not be able to keep pace with assays of the promoter regions.

The 5′ SMART RACE experiments performed to identify the predicted first exon of each gene, were also a slowing point in this high throughput pipeline. Confirming the first exon is going to be critical for proving which gene these promoters operate on, however, it may take more time. Trying different RACE kits or alternatively amplifying RT-PCR products with gene specific primers might yield better results. Of the four promoters tested only one, dk3, was located adjacent to the first exon of a gene (*DKFZp564A1164*). The other 3 promoters are 6 to 10 kb away from the known transcribed sequences of the genes they are predicted to operate on.

The failure to RACE*NPHS1* and *HSPOX1* could also be taken, together with reporter results, to indicate that FirstEF failed to find either gene's promoter and that the prediction of upstream exons may be incorrect. In this case it is clear that FirstEF did fail to predict what appear to be the most commonly used first exons for both *NPHS1* and *HSPOX1*. One of the purposes of this study was to provide data to test FirstEF predictions and feedback the results to FirstEF's creators. Because FirstEF is a relatively new program, such feedback will be helpful in refining its prediction algorithms.

The n2 construct, which was predicted to operate in one orientation and to provide a potential upstream promoter for *NPHS1* gene, was found to be a strong bi-directional promoter. The closest gene that the reverse orientation of n2 (n2r) could be operating on is *APLP1,* 8 Kb away. However, it is more likely that n2r could be an internal alternative promoter for *DKFZp564A1164,* and may potentially define alternative start sites for both *NPHS1* and *DKFZp564A1164*. Further experiments are necessary in this region to confirm which genes these promoters are operating on.

Conclusion

For this study, a high throughput method for identifying and testing regulatory elements was examined. In addition, the validity of promoters predicted by FirstEF was tested. It was found that by combining computer based promoter and first exon predictions from FirstEF (Davuluri *et al.*, 2001) with PCR-based cloning to generate luciferase reporter constructs, and by testing reporter activity in cultured mammalian cells plated in a 96 well format one could identify promoter activity in a relatively high throughput manner.

The data generated in this study suggest that FirstEF predictions are sometimes incorrect. Therefore, having a strategy for defining which FirstEF predicted promoters to test first may accelerate the process. Initially testing promoters that are at a confirmed transcription start site for a gene, at a possible alternate transcription start site or in a region of conserved sequence would be the best candidates, while promoters predicted in gene desert regions may not be as easy to confirm.

The luciferase assay lent itself very well to the high throughput search, however the subcloning did not always go smoothly. The numerous steps that this traditional subcloning method requires were time consuming and increased the opportunities for errors. A faster method that skips many of the traditional subcloning steps, such as the Creator™ system by Clontech is currently being investigated by our lab.

The development and testing of substantially larger enhancer/ silencer regulatory elements may not be possible at this time using these high throughput methods. These regulatory elements are generally GC rich making them more difficult to PCR and subclone. Additionally, confirming upstream untranslated first exons was not possible within this time scale using the SMART RACE protocol. It will be necessary to further explore the limitations within these procedures in order to confirm these and future regulatory elements. Alterations and modifications to these protocols, as well as investigating new techniques may be necessary.

REFERENCES

Adam G. I., Miller S. J., Ulleras E., and Franklin G. C. (1996). Cell-type-specific modulation of PDGF-B regulatory elements via viral enhancer competition: a caveat for the use of reference plasmids in transient transfection assays. *Gene* **178:** 25-29.

Asnagli H., Afkarian M., and Murphy K. M. (2002). Identification of an Alternative GATA-3 Promoter Directing Tissue- Specific Gene Expression in Mouse and Human. *J Immunol* **168:** 4268-71.

Belsham D. D., and Mellon P. L. (2000). Transcription factors Oct-1 and C/EBPbeta (CCAAT/enhancer-binding protein-beta) are involved in the glutamate/nitric oxide/cyclic-guanosine 5'-monophosphate-mediated repression of mediated repression of gonadotropin-releasing hormone gene expression. *Mol. Endocrinol.* **14:**212- 228.

Bray N., Dubchak, I. and Pachter, L (2003). AVID: A Global Alignment Program. *Genome Research* **13:** 97.

Chakravarti A. (2002). A compelling genetic hypothesis for a complex disease: PRODH2/DGCR6 variation leads to schizophrenia susceptibility. *Proc. Natl. Acad. Sci. USA*, **99:**4755 -4756.

Davuluri R. V., Grosse I., and Zhang M. Q. (2001). Computational identification of promoters and first exons in the human genome. *Nat Genet* **29:**412- 7.

Dehal P., Predki P., Olsen A. S., Kobayashi A., Folta P., Lucas S., Land M., Terry A., Ecale Zhou C. L., Rash S., Zhang Q., Gordon L., Kim J., Elkin C., Pollard M. J., Richardson P., Rokhsar D., Uberbacher E., Hawkins T., Branscomb E., and Stubbs L. (2001). Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293:** 104-11.

Dubchak I., Brudno M., Loots G. G., Pachter L., Mayor C., Rubin E. M., and Frazer K. A. (2000). Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res* **10:**1304 -6.

Goodman B., Rutberg J., Lin W., Pulver A., and Thomas G. (2000). Hyperprolinaemia in patients with deletion (22)(q11.2) syndrome. *J Inherit Metab Dis* **23:** 847-848.

Hallman N., Hjelt L., and Ahvenainen E. K. (1956). Nephrotic syndrome in newborn and young infants. *Ann Paediat Fenn* **2:** 227-241.

Hardison R. C., Oeltjen J., and Miller W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* **7:** 959-966.

Havik B., Ragnhildstveit E., Lorens J. B., Saelemyr K., Fauske O., Knudsen L. K., and Fjose A. (1999). A novel paired domain DNA recognition motif can mediate Pax2 repression of gene transcription. *Biochem Biophys Res Commun.* **266:** 532-541.

Hood L., Koop B. F., Rowen L., and Wang K. (1993). Human and mouse T-cell-receptor loci: the importance of comparative large-scale DNA sequence analyses. *Cold Spring Harb Symp Quant Biol* **58:** 339-348.

Ihalmo P., Palmen T., Ahola H., Valtonen E., and Holthofer H. (2003). Filtrin is a novel member of nephrin-like proteins. *Biochem Biophys* **300:** 364-370.

Kemp D. M., Lin J. C., Ubeda M., and Habener J. F. (2002). NRSF/REST confers transcriptional repression of the GPR10 gene via a putative NRSE/RE-1 located in the 5' promoter region. *FEBS Lett* **531:** 193-8.

Kestila M., Lenkkeri U., Mannikko M., Lamerdin J., McCready P., Putaala H., Ruotsalainen V., Morita T., Nissinen M., Herva R., Kashtan C. E., Peltonen L., Holmberg C., Olsen A., and Tryggvason K. (1998). Positionally cloned gene for a novel glomerular protein--nephrin--is mutated in congenital nephrotic syndrome. *Mol Cell* **1:** 575-82.

Kim J., Kollhoff A., Bergmann A., and Stubbs L. (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. *Hum Mol Genet.* **12:**233 -245.

Kraner S. D., Chong J. A., H.J. T., and Mandel G. (1992). Silencing the type II sodium channel gene: a model for neural-specific gene regulation. *Neuron* **1:** 37-44.

Lenkkeri U., Mannikko M., McCready P., Lamerdin J., Gribouval O., Niaudet P. M., Antignac C. K., Kashtan C. E., Homberg C., Olsen A., Kestila M., and Tryggvason K. (1999). Structure of the gene for congenital nephrotic syndrome of the finnish type (NPHS1) and characterization of mutations. *Am J Hum Genet* **64:** 51-61.

Liu M., Leibowitz J. L., Clark D. A., Mendicino M., Q. N., Ding J. W., D'Abreo C., Fung L., Marsden P. A., and Levy G. A. (2003). Gene transcription of fgl2 in endothelial cells is controlled by Ets-1 and Oct-1 and requires the presence of both Sp1 and Sp3. *Eur J Biochem.* **270:** 2274-2286.

Loots G. G., Locksley R. M., Blankespoor C. M., Wang Z. E., Miller W., Rubin E. M., and Frazer K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136-40.

Loots G. G., Ovcharenko, I., Pachter, L., Dubchak, I., Rubin, E. (2002). rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12:**832 -9.

Mather J. P., and Roberts P. E. (1998). "Introduction to cell and tissue culture: theory and technique," Plenum Press, New York.

Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S., and Dubchak I. (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:**1046- 7.

Moeller M. J., Kovari I. A., and Holzman L. B. (2000). Evaluation of a new tool for exploring podocyte biology: mouse Nphs1 5' flanking region drives LacZ expression in podocytes. *J Am Soc Nephrol* **11:** 2306-14.

Parsons S. J., Rhodes S. A., Conner H.E., Rees S., Brown J., Giles H. (2000). Use of a dual firefly and Renilla luciferase reporter gene assay to simultaneously determine drug selectivity at human corticotrophin releasing hormone 1 and 2 receptors. *Analytical Biochemistry* **281:** 187-192.

Praz V., Perier R., Bonnard C., and Bucher P. (2002). The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res* **30:** 322-324.

Schwartz S., Zhang Z., Frazer K. A., Smit A., Riemer C., Bouck J., Gibbs R., Hardison R., and Miller W. (2000). PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:**577- 86.

Shen Z. J., Nakamoto T., Tsuji K., Nifuji A., Miyazono K., Komori T., Hirai H., and Noda M. (2002). Negative regulation of bone morphogenetic protein/Smad signaling by Cas-interacting zinc finger protein in osteoblasts. *J Biol Chem.* **277:** 29840-6.

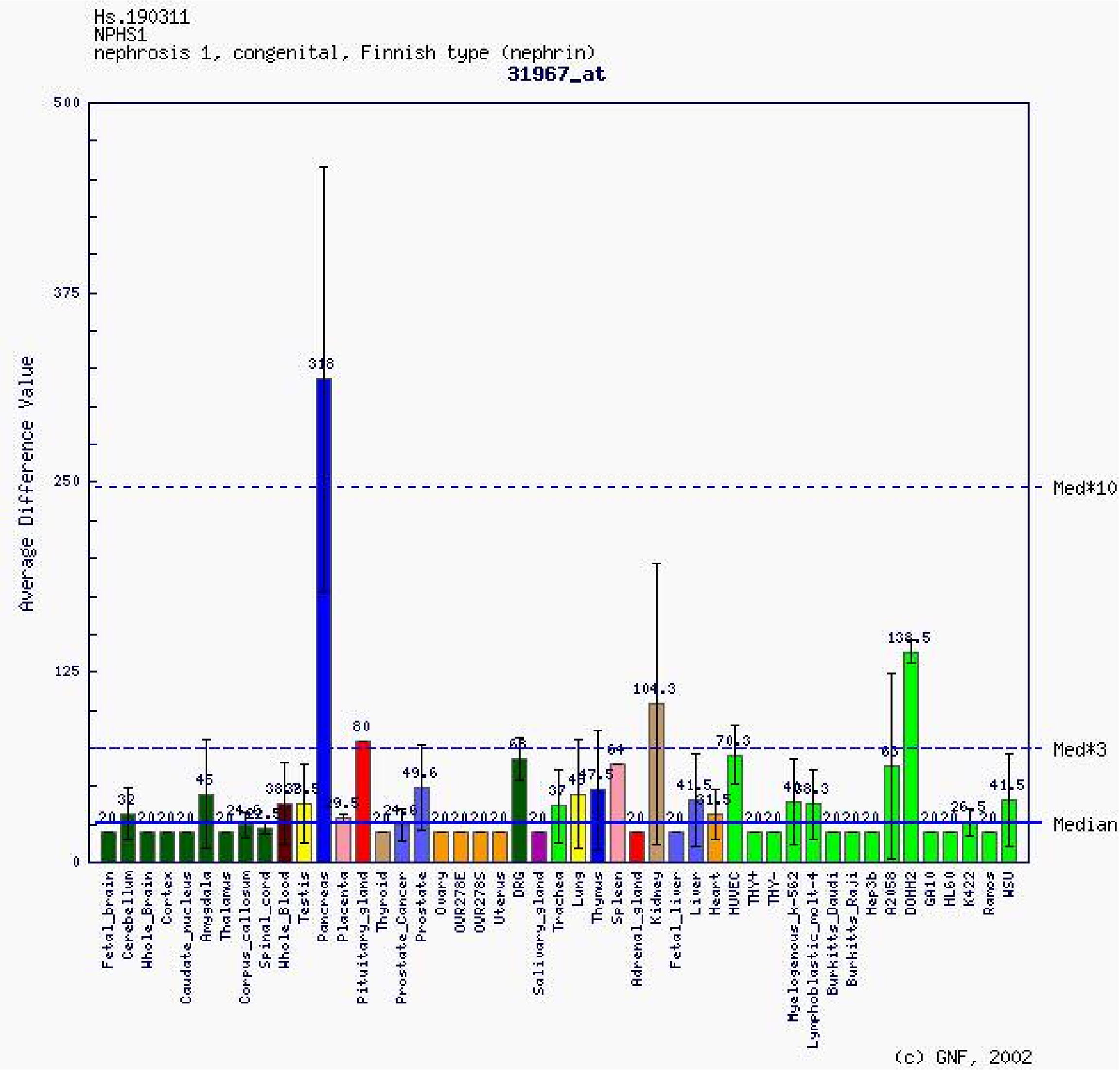
Sherf B. A., Navarro S. L., Hannah R. R., Wood K. V. (1996). Dual- luciferase reporter assay: an advanced co-reporter technology integrating firefly and renilla luciferase assays. *In* "Promega Notes Magazine", pp. 02.

Wiemann S., Weil B., Wellenreuther R., Gassenhuber J., Glassl S., Ansorge W., Bocher M., Blocker H., Bauersachs S., Blum H., Lauber J., Dusterhoft A., Beyer A., Kohrer K., Strack N., Mewes H. W., Ottenwalder B., Obermaier B., Tampe J., Heubner D., Wambutt R., Korn B., Klein M., and Poustka A. (2001). Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res* **11:** 422-35.

Zhou Z.-Q., and Walter C. A. (1998). Cloning and characterization of the promoter of baboon XRCC1, a gene involved in DNA strand-break repair. *Somat Cell Mol Genet.* **24:**23- 39.

APPENDIX

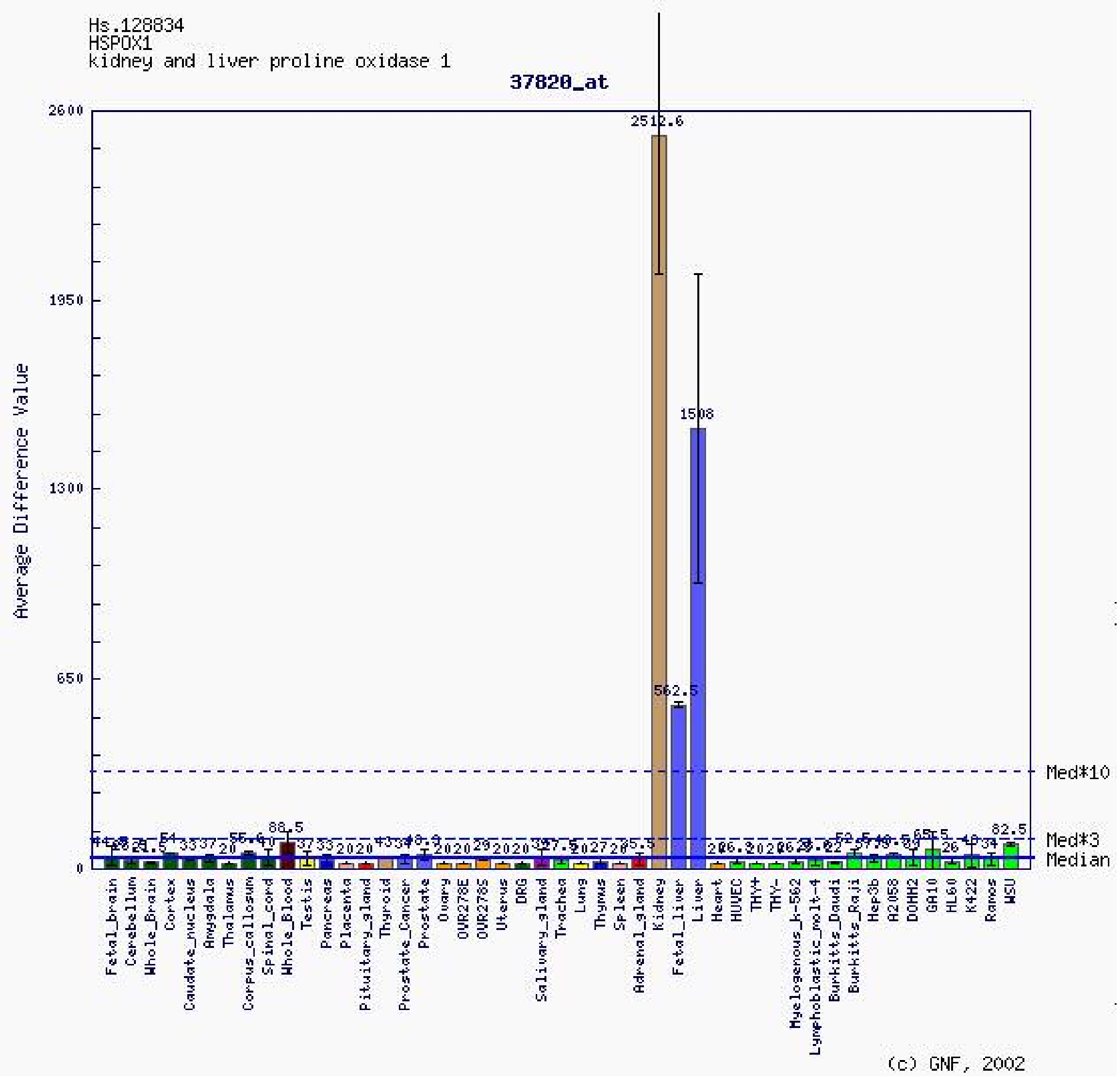
Figure 1A: Microarray expression data



Microarray expression data from Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi) microarray database for *NPHS1* showing highest expression in the pancreas.

APPENDIX

Figure 2A: Microarray expression data



Microarray expression data from Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi) microarray database for *HSPOX1* showing highest expression in the kidney.

APPENDIX

Table 1A: Primers for cDNA analysis.

Gene name	Forward primer (exon)	Reverse primer (exon)
NPHS1	GAG GAC CGA GTC AGG AAC GAA (26)	CTG CAC TTC ATC GTA GAG GGG T (28)
DKFZp564A1164	AGC AAA AGA ACC TGA TGC GAA TC (13)	TTG ATG TAG CTG GTG AAA GCT CG (15)
HSPOX1	CCA TGA GGA ARC TGT TCG CC (9)	TGC TAG TGG GGT ATC CTT C (11)
β−actin	GCG GGA AAT CGT GCG TGA CAT T	GAT GGA GTT GAA GGT AGT TTC GTG

Table 2A: Tissue array probes.

Gene	***HSPOX1***
accession number	NM_021232, mRNA
forward	GGG CAG TTG GTG AAC TTG CT
reverse compliment	TCA GCT CTC CTG TGC CCT TA
reverse w/ **t7**	**TAATACGACTCACTATAGGG**TCA GCT CTC CTG TGC CCT TA
gene	***NPHS1***
accession number	NM_004646
forward	GAG GAG GTG TCT TAT TCC CG
reverse compliment	TCC AGA GTG TCC AAG TCT CC
reverse w/ **t7**	**TAATACGACTCACTATAGGG**TCC AGA GTG TCC AAG TCT CC
gene	***DKFZP564A1164***
accession number	NM_032123
forward	ACT ACA AGG TCC GAG GAG TC
reverse compliment	TGC CCT GGC TCT GTA AAG TC
reverse w/ **t7**	**TAATACGACTCACTATAGGG**TGC CCT GGC TCT GTA AAG TC

APPENDIX

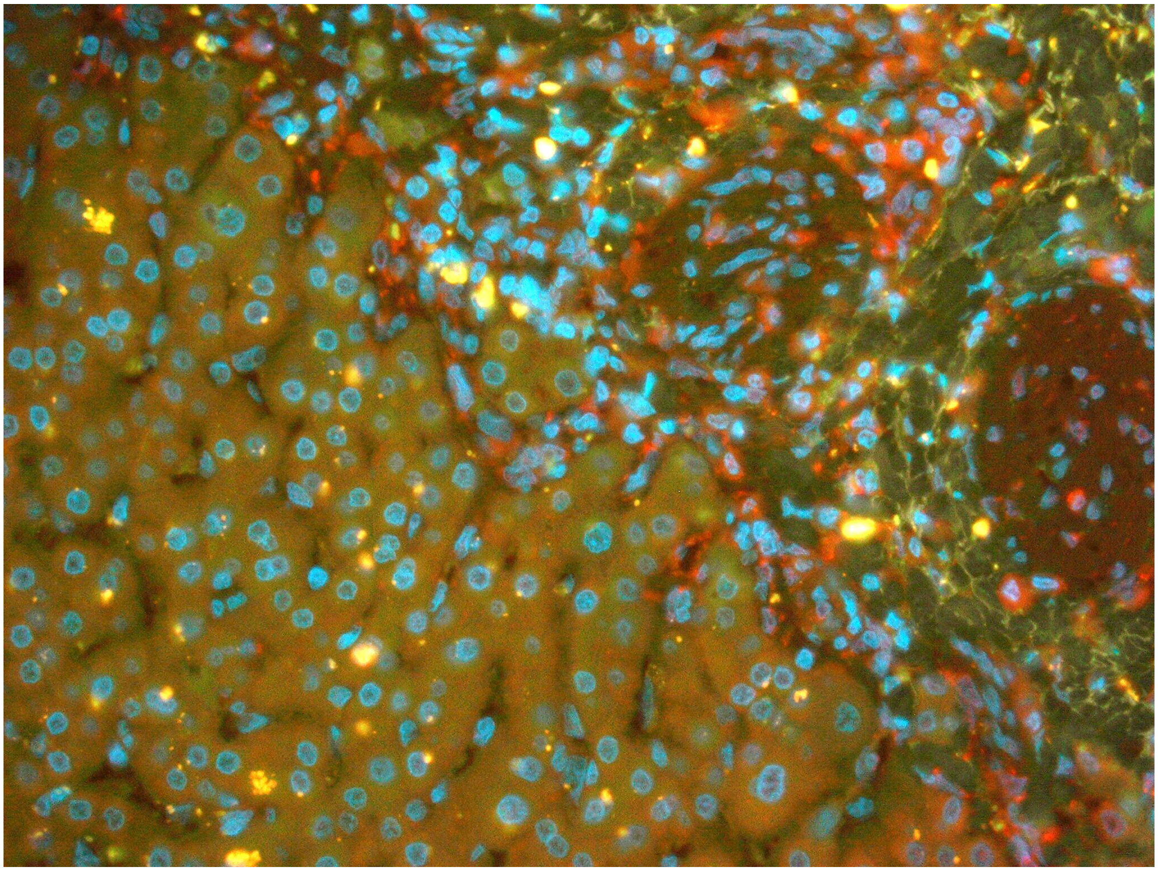
Table 3A: Tissue hybridization results.

	NPHS1	***HSPOX1***	***DKFZp564A1164***
Lung	-	-	+ / -
Skin	+	+ / -	+ exterior
Muscle, skeletal	-	-	-
Heart, muscle	-	-	+ / -
Stomach	-	++	++
Esophagus	+/-	+ / -	+ / -
Small intestine	+	+ / -	+
Colon	-	-	+ / -
Liver	-	+ / -	+
Spleen	-	-	-
Pancreas	+	-	-
Salivary gland	-	+	+ / -
Pituitary gland	-	-	-
Adrenal gland	-	-	-
Thyroid gland	-	-	- (+ support tissue)
Parathyroid gland	-	-	-
Thymus gland	+/-	+ / -	++
Tonsil	+	+	+
Bone marrow	-	-	-
Breast	- (+ in gland)	- (+ in gland)	- (+ in gland)
Uterus	++	+	++
Cervix	+/-	-	+/- endothelial
Ovary	++	++	++
Kidney	+ (tubules only)	+ / -	+ (tubules only)
Prostate gland	++	+	++
Testis	++	++	++
Omentum	-	+ / -	+
Peripheral nerve	+	+ / -	+
Cerebral cortex	++	++	+
Cerebellum	++	+	+ perkingi

Table 3A : – negative; +/- weak positive; + positive; ++ strong positive

APPENDIX

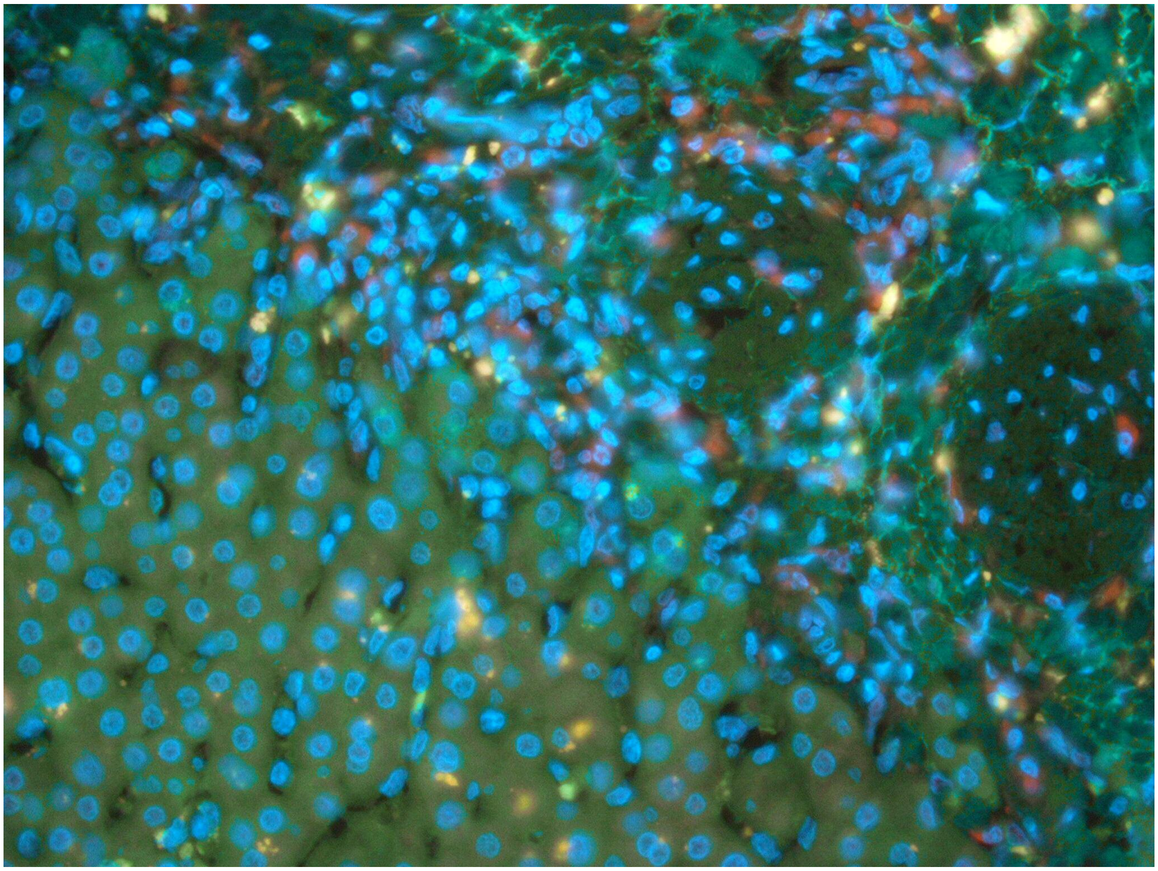
Figure 3A: Liver tissue slide



Liver tissue: *DKFZp564A1164* t7 mRNA probe labeled with dig and hybridized to a MaxArray normal human tissue slide (Zymed Laboratories, Inc.) red color indicates positive hybridization

APPENDIX

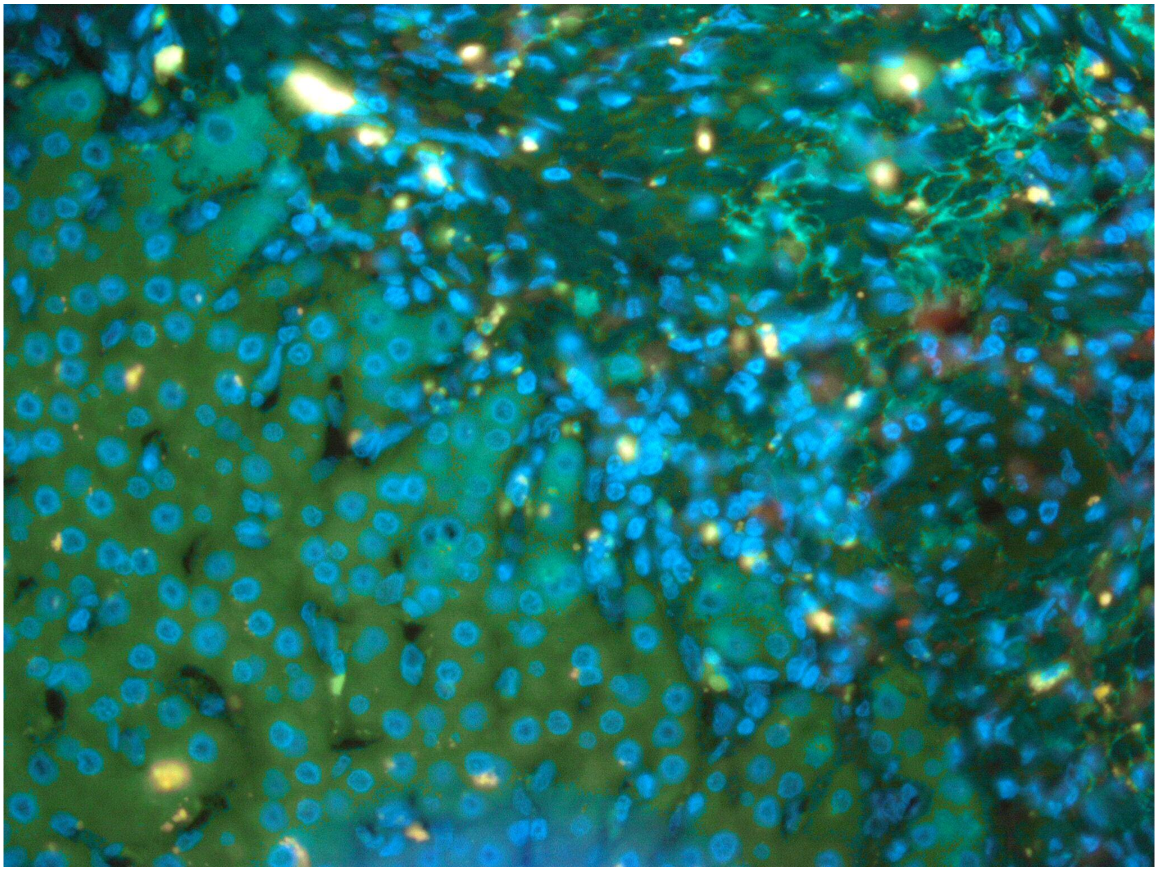
Figure 4A: Liver tissue slide



Liver tissue: *HSPOX1* t7 mRNA probe labeled with dig and hybridized to a MaxArray normal human tissue slide (Zymed Laboratories, Inc.) red color indicates positive hybridization

APPENDIX

Figure 5A: Liver tissue slide



Liver tissue: *NPHS1* t7 mRNA probe labeled with dig and hybridized to a MaxArray normal human tissue slide (Zymed Laboratories, Inc.) red color indicates positive hybridization.

.

APPENDIX

Table 4A: Primers for PCR of promoter or promoter + enhancer constructs.

Name	**Forward** (lowercase letters are restriction enzyme sequence)	**Reverse**(lowercase letters are restriction enzyme sequence)
n1	gga aga tct CTG CAG GCA AAG CCG GAG CC	cgg ggt acc ccg AGG TTT GGA GGT CTC
n2	gga aga tct CTG CAG GCA AAG CCG GAG CC	cgg ggt acc ccA AAG GCT GTA ACA AAG CC
n3	gga aga tct tcc ACT CTC TCC CTT CCC TCC	cga cgc gtc gTT CTC GCT AGT GAA GAG GCA
n4	gga aga tct tcc ACT CTC TCC CTT CCC TCC	cga cgc gtc gTC TCG AAC TCC TGA TCT TAG
n2r	cgg ggt acc ccT GCA GGC AAA GCC GGA GCC	gga aga tct AAA GGC TGT AAC AAA GCC
n3r	cga cgc gtc Gtc TTC CAC TCT CTC CCT TCC	gga aga tct TCT CGC TAG TGA AGA GGC A
dk1	cgg ggt acc ccg AAG GAC GCT CCT GGC GGC	gga aga tct tcc AAG GCT GGA CAG CTC AGC
dk2	cgg ggt acc ccg TGT GAG AGG GCC CCA GGT	gga aga tct tcc AAG GCT GGA CAG CTC AGC
dk3	cga cgc gtc gaA TTG AGC TGG GGG CGC CCA	gga aga tct tcc GGG GCA GCA GGG CTG AGC
dk4	cga cgc gtc gaA ATC CTC CTG GGC CTG TG	gga aga tct tcc GGG GCA GCA GGG CTG AGC
dk1r	gga aga tct tcc AAG GAC GCT CCT GGC GGC	cgg ggt acc ccg AAG GCT GGA CAG CTC AGC
dk3r	gga aga tct tcc TTG AGC TGG GGG CGC CCA	cga cgc gtc gaG GGG CAG CAG CGG CTG AGC
dk3 large	gga aga tct tcc ACT CTC TCC CTT CCC TCC	cga cgc gtc gaG GGG CAG CAG CGG CTG AGC
n3 large	gga aga tct tcc GGG GCA GCA GGG CTG AGC	cga cgc gtc Gtc TTC CAC TCT CTC CCT TCC
dk3 small	gga aga tct tcc TTG AGC TGG GGG CGC CCA	cga cgc gtc gTT CTC GCT AGT GAA GAG GCA

APPENDIX

Table 5A: Primers for 5’ SMART RACE of *NPHS1* and *HSPOX1*

NPHS1 Race primers

Name	**Sequence**	**Size**
rn1	GGATGGAGAGGATCACTCTGGGAGACACGA	30 bp
rn2	CCTGAAAACCTGACGGTGGTGGAGGGGGCC	30 bp
rn3	CGGAGTATGAGTGCCAGGTCGGCCGCTCTG	30 bp

HSPOX1 RACE primers

Name	**Sequence**	**Size**
rh1	GGGAACAGAGCACGTAACAGGTCCGGAGC	29 bp
rh2	CTCACCAGCCACAAACTGCCCATAGACGG	29 bp
rh3	ATAGCACCGAGGTTCCCCTCATACCACGCC	30 bp

APPENDIX

Table 6A: Transcription factor binding sites (TFBS) found by rVISTA

Promoter/ Enhancer	**TFBS**	**Number of Hits**
n1 = 11 conserved TFBS enhancer region	AP2ALPHA	2
	CAP	5
	GATA	1
	TEF1_Q6	1
	GEN_INI_B	1
	HOXA4_Q2	1
n2 = 12 aligned TFBS promoter region	CAP	8
	CETS1P54	1
	ZIC3	1
	CDXA	1
	MZF1	1
n3 = 9 conserved TFBS promoter region	CAP	2
	STAT	2
	CETS1P54	1
	PAX2	2
	MYB_Q6	1
	SRY	1
n4 = 60 conserved TFBS enhancer region	MYB_Q6	1
	CAP	20
	CDXA	1
	STAT	5
	PAX2	6
	PAX4	1
	HOXA4_Q2	2
	TEF1_Q2	1
	GEN_INI_B	4
	GATA	2
	CEBP	1
	TCF4_Q5	1
	CETS1P54	2
	NFAT_Q6	1
	YY1	2
	PEA3_Q6	1
	AP2ALPHA	1
	SPZ1	1
	DBP_Q6	1
	EN1	1
	GR_Q6	1
	PU1_Q6	1
	NKX62_Q2	1
	OCT1	1
	CIZ	1

APPENDIX

Table 6A: **Transcription factor binding sites (TFBS) found by rVISTA**

Promoter/ Enhancer	**TFBS**	**Number of Hits**
dk1 = 2 conserved TFBS promoter region	CAP	1
	ZP1	1
dk3 = 4 conserved TFBS promoter region	STAT	1
	PAX2	2
	CAP	1
	MYB_Q6	1
	SRY	1
dk4 = 20 conserved TFBS enhancer region	PAX2	2
	CIZ	1
	STAT	4
	LPOLYA_B	1
	CDXA	1
	GATA	2
	CAP	6
	HSF1	1
	AP2ALPHA	1
	CETS1P54	1